

The Nu Expression for Probabilistic Data Integration

Evgenia I. Polyakova · Andre G. Journel

Received: 16 October 2006 / Accepted: 16 April 2007 / Published online: 12 October 2007
© International Association for Mathematical Geology 2007

Abstract The general problem of data integration is expressed as that of combining probability distributions conditioned to each individual datum or data event into a posterior probability for the unknown conditioned jointly to all data. Any such combination of information requires taking into account data interaction for the specific event being assessed. The nu expression provides an exact analytical representation of such a combination. This representation allows a clear and useful separation of the two components of any data integration algorithm: individual data information content and data interaction, the latter being different from data dependence. Any estimation workflow that fails to address data interaction is not only suboptimal, but may result in severe bias. The nu expression reduces the possibly very complex joint data interaction to a single multiplicative correction parameter ν_0 , difficult to evaluate but whose exact analytical expression is given; availability of such an expression provides avenues for its determination or approximation. The case $\nu_0 = 1$ is more comprehensive than data conditional independence; it delivers a preliminary robust approximation in presence of actual data interaction. An experiment where the exact results are known allows the results of the $\nu_0 = 1$ approximation to be checked against the traditional estimators based on assumption of data independence.

Keywords Data integration · Data interaction vs. dependence · Updating probabilities · Conditional independence

E.I. Polyakova (✉) · A.G. Journel
Department of Geological and Environmental Sciences, Stanford University, Palo Alto, CA 94305,
USA
e-mail: jencya@pangea.stanford.edu

A.G. Journel
e-mail: journal@pangea.stanford.edu

Introduction

Combining information from different sources is a challenging yet recurrent task in many scientific fields. Statisticians view this problem as combining prior and preposterior probabilities into a posterior probability (e.g., Genest and Zidek 1986; Caers and Hoffman 2006). Most often some form of data conditional independence is assumed to obtain the posterior probabilities, which may lead to incorrect and possibly nonconservative conclusions, if any datum transforms the information brought by the other data. In geology, there are rarely satisfying grounds to assume data independence or conditional independence. Different data relate to events that are often associated to a common geological origin. From one location to another nearby location geological structures are related one to another, leaving us with the challenge of building posterior probabilities that do not start by assuming any form of independence.

The assumption of data independence is routinely used in, for example, linear regression theory, thus the origin of the name “independent variables” (Galton 1894). The independence assumption was overcome, only to a degree, by introducing the variogram/covariance concept which incorporates linear dependence between any two data values or events, for example, as observed at two different locations in space. However, complex geological patterns whose description involves multiple locations in space were beyond the reach of traditional two-point geostatistics until the recent introduction of multiple-point geostatistics (Strebelle 2001).

Any individual datum information can be coded into a conditional probability; this raises the question of how to combine such different probabilities. One simplistic solution for combination is to consider weighted linear averages of prior probabilities to obtain the final posterior conditional probability, as done in indicator kriging (Journel 1983). For example, assume we are trying to find the probability $P(A|B, C)$ conditioned to the two data events B and C . One possible solution is

$$P(A|B, C) = \lambda_1 P(A|B) + \lambda_2 P(A|C).$$

For such a linear combination to take a value between $[0, 1]$, one typically restricts the weights λ to be positive and sum to one which entails convexity of the result: the combined probability is then bounded by the two prior probabilities $P(A|B)$ and $P(A|C)$. Such convexity is undesirable because it precludes the possibility of compounding, say, the high probability $P(A|B)$ with the concordant information carried by $P(A|C)$ into a combined probability $P(A|B, C)$, higher than either $P(A|B)$ or $P(A|C)$.

The multiplicative tau model, initially introduced by Bordley (1982), has been extensively used as a heuristic for differently weighting prior probability distributions (Benediktsson and Swain 1992). This model has been further developed by Journel (2002). It allowed an elegant solution to the problem of combining prior probabilities with the resulting posterior probability being always licit and accounting, not only for the individual information brought by the different data sources, but also for data interaction through the tau weights, see later relation (8).

This tau model remained a model, that is an approximation of the exact conditional probability given all data events taken jointly, until the seminal contribution of Krishnan (2005). Krishnan established the closure of the tau model by giving the exact

analytical expression of the tau weights. When these values for the tau weights are used, the tau model (8) is not a model any more; instead it is the exact representation of the combined probability. From this exact expression, avenues for approximating the tau weights have been proposed (Krishnan 2005).

In this paper, an alternative nu representation is proposed. The nu expression is derived directly, independently of the original tau expression, and is shown to be strictly equivalent to the latter. The nu expression is appealing because its derivation is more straightforward and it puts forward a single correction parameter, v_0 , accounting for global data interaction and independent of the sequence along which the data are considered.

The Nu Representation

Consider an unknown event A informed by n diverse data events $D_1, \dots, D_n$. These data have been evaluated for their individual information content related to unknown event A through the elementary probabilities $P(A|D_i)$.¹ The challenge is then to recombine the prior probability $P(A)$ and the n single-event conditional probabilities $P(A|D_i)$ into the posterior probability $P(A|D_1, \dots, D_n)$ while accounting for the interaction among data. The nu representation provides a simple yet exact solution for such recombination.

To establish the nu expression, first define the following distances to the unknown event (A) prior to and after observing any single data event $x_0 = \frac{1-P(A)}{P(A)} = \frac{P(\tilde{A})}{P(A)} \in [0, \infty]$; prior distance to A occurring $x_i = \frac{1-P(A|D_i)}{P(A|D_i)} = \frac{P(\tilde{A}|D_i)}{P(A|D_i)} \in [0, \infty]$; updated distance knowing datum D_i , equals to zero if $P(A|D_i) = 1$, equals to infinity if $P(A|D_i) = 0$. The updated distance knowing jointly the n data is then $x = \frac{1-P(A|D_1, \dots, D_n)}{P(A|D_1, \dots, D_n)} = \frac{P(\tilde{A}|D_1, \dots, D_n)}{P(A|D_1, \dots, D_n)} \in [0, \infty]$.

The one to one relation between the updated distance and the fully conditioned probability is simply

$$P(A = a|D_i = d_i, i = 1, \dots, n) = \frac{1}{1 + x} \in [0, 1] \quad (1)$$

where $\tilde{A} = \text{non}A$. If event $A = a$ is a nonbinary random variable, then $\text{non}A$ denotes any A -outcome different from a .

Note that a distance is but the inverse of an odds ratio (Fleiss 1981), e.g., the prior odds of event A occurring is

$$\frac{P(A)}{1 - P(A)} = \frac{1}{x_0}.$$

Next, consider the following exact decomposition of the joint conditional probability

¹Throughout this paper, the short notation $P(A|D_i)$ will be used for the (a, d_i) values-specific expression $P(A = a|D_i = d)$.

$$\begin{aligned}
P(A|D_1, \dots, D_n) &= \frac{P(A, D_1, \dots, D_n)}{P(D_1, \dots, D_n)} \\
&= \frac{P(A)P(D_1|A)P(D_2|A, D_1) \cdots P(D_n|A, D_1, \dots, D_n)}{P(D_1, \dots, D_{n-1})} \\
&= P(A) \frac{\prod_{i=1}^n P(D_i|A, \bar{D}_{i-1})}{P(D_1, \dots, D_n)} \quad (2)
\end{aligned}$$

where $\bar{D}_{i-1} = \{D_1 = d_1, \dots, D_{i-1} = d_{i-1}\}$, i.e., the set of all data up to the $(i-1)$ th datum $D_{i-1} = d_{i-1}$.

The notation $\bar{D}_{i-1}$ implies a specific sequence of n data, with D_1 being the first datum considered and D_n being the last. There are $n!$ such data sequences.

Then, rewrite the fully updated distance x as the product series

$$x = \frac{P(\tilde{A}) \frac{\prod_{i=1}^n P(D_i|\tilde{A}, \bar{D}_{i-1})}{P(D_1, \dots, D_n)}}{P(A) \frac{\prod_{i=1}^n P(D_i|A, \bar{D}_{i-1})}{P(D_1, \dots, D_n)}} = x_0 \prod_{i=1}^n \frac{P(D_i|\tilde{A}, \bar{D}_{i-1})}{P(D_i|A, \bar{D}_{i-1})}. \quad (3)$$

The ratio $\frac{P(D_i|\tilde{A}, \bar{D}_{i-1})}{P(D_i|A, \bar{D}_{i-1})}$ measures how the likelihood of observing the datum $D_i = d_i$ changes whether A or *nonA* is present, once the previous $(i-1)$ data values $\bar{D}_{i-1} = \bar{d}_{i-1}$ have been observed. This ratio is not yet a measure of interaction between the D_i -datum and the previous data considered $\bar{D}_{i-1}$. A measure of such interaction would be obtained by comparing the previous ratio to the one ignoring $\bar{D}_{i-1}$, that is to $\frac{P(D_i|\tilde{A})}{P(D_i|A)}$. Consider then the nu parameter defined as the ratio of ratios

$$\nu_i = \frac{\frac{P(D_i|\tilde{A}, \bar{D}_{i-1})}{P(D_i|A, \bar{D}_{i-1})}}{\frac{P(D_i|\tilde{A})}{P(D_i|A)}} \geq 0, \quad \text{with } \nu_1 = 1. \quad (4)$$

The denominator of (4) can be rewritten with Bayes inversion

$$\frac{P(D_i|\tilde{A})}{P(D_i|A)} = \frac{P(D_i, \tilde{A})}{P(\tilde{A})} \frac{P(A)}{P(D_i, A)} = \frac{P(\tilde{A}|D_i)P(D_i)}{P(\tilde{A})} \frac{P(A)}{P(A|D_i)P(D_i)} = \frac{x_i}{x_0}.$$

Hence, (3) can be rewritten as the *exact* nu equation

$$\frac{x}{x_0} = \prod_{i=1}^n \nu_i \frac{x_i}{x_0} = \nu_0 \prod_{i=1}^n \frac{x_i}{x_0}, \quad \text{where } \nu_0 = \prod_{i=1}^n \nu_i \geq 0. \quad (5)$$

Equations (3) and (5) are two exactly equivalent expressions for the conditional distance x , but (5) is written as the product of the more easily accessible elementary distances x_i with parameters (the ν_i and ν_0) interpreted hereafter as measures of data interaction. It is hoped that these ν_i -parameters [see (4)] being ratios of ratios are reasonably stable, hence easier to infer from proxy training data than the direct likelihood ratios of (3).

Remarks and Interpretation

- Recall once again that all probabilities in (1) to (5), and therefore, the (v_0, v_i) parameters are data values-dependent (the d'_i 's). In addition, and most critically, they are also dependent on the specific outcome $A = a$ being discriminated from $A = \text{nona}$. Note that this discrimination is between a and nona , *not* between any two outcomes $a_k, a_{k'}$ of a variable A with more than two outcomes $\{a_k, k = 1, \dots, K > 2\}$.
- The denominator of the v_i -expression (4) measures how datum $D_i = d_i$ discriminates the outcome $A = a$ from nona . The numerator measures the same, but in the presence of the previous data $\overline{D}_{i-1} = \overline{\mathbf{d}}_{i-1} = \{D_1 = d_1, \dots, D_{i-1} = d_{i-1}\}$. Thus, the ratio (of ratios) v_i indicates how the discrimination power of $D_i = d_i$ is changed by knowledge of the previous data $\overline{D}_{i-1} = \overline{\mathbf{d}}_{i-1}$. Note that the n parameters v_i are *data sequence-dependent*, as opposed to the single parameter v_0 in (5).
- Equation (4) of v_i is symmetric in D_i and $\Delta_j = \overline{D}_{i-1}$: one can exchange D_i for Δ_j , the expression of v_i would remain unchanged.
- If $v_i = 1$, the ability of the datum (or data event) $D_i = d_i$ to discriminate a from nona is unchanged by knowledge of the previous $(i - 1)$ data events $\overline{\mathbf{d}}_{i-1} = \{D_1 = d_1, \dots, D_{i-1} = d_{i-1}\}$. Therefore, we can consider $v_i = 1$ as the case of “noninteraction” of the two data events D_i and $\overline{D}_{i-1}$ when it comes to discriminating a from nona . The deviation $|1 - v_i|$ is thus a measure of data interaction. That measure is $(a, d_j, j = 1, \dots, i)$ values-dependent, although at times, we will use the short language “ $D_i, \overline{D}_{i-1}$ data interaction”.

Note that noninteraction is *not* independence, if only because data independence, say between two random variables D_1 and D_2 , does not involve any third variable A . The term “data redundancy” is sometimes used instead of “data interaction”; this is unfortunate because redundancy implies some overlap of the two data D_1 and D_2 and redundancy of information being a state one should avoid or, at least, correct. We prefer the more neutral term “data interaction” which means only that one datum modifies the information brought by the other whenever $v_i \neq 1$, i.e., $|1 - v_i| > 0$.

- $v_i = 1$ in (4) requires that the two ratios $\frac{P(D_i|\tilde{A}, \overline{D}_{i-1})}{P(D_i|A, \overline{D}_{i-1})}$ and $\frac{P(D_i|\tilde{A})}{P(D_i|A)}$ be equal to each other. One sufficient (but not necessary) condition for $v_i = 1$ is conditional independence of the two data D_i and $\overline{D}_{i-1}$ given $A = a$ and given $A = \text{nona}$, that is

$$\begin{aligned} P(D_i|A = a, \overline{D}_{i-1}) &= P(D_i|A = a) \quad \text{and} \\ P(D_i|A = \text{nona}, \overline{D}_{i-1}) &= P(D_i|A = \text{nona}). \end{aligned}$$

Note that the traditional conditional independence of D_i and $\overline{D}_{i-1}$ given only $A = a$ does not suffice to ensure $v_i = 1$. See example in the Appendix. Thus data independence, data conditional independence and data non-interaction ($v_i = 1$) are different states.

- If $v_i > 1$ in (5), the relative distance x_i/x_0 is increased.

1. If $x_i > x_0$, that is, if datum D_i by itself increases the distance to event $A = a$ occurring as compared to the prior distance x_0 , then the interaction factor $v_i > 1$ makes that increase even greater.
 2. Similarly, if $x_i < x_0$, that is, if datum D_i by itself decreases the distance to event $A = a$ occurring, then the interaction factor $v_i > 1$ reduces that decrease.
- If $v_i < 1$, the previous conclusions are reversed.
 - The greater the data interaction as measured by the deviation $|1 - v_i|$, the more critical it is to consider the data $\overline{D}_{i-1}$ and D_i jointly because they interact one on the other as to evaluating the outcome $A = a$. Thus, one can also read $|1 - v_i|$ as a measure of the information modification brought by jointly considering the two data $\overline{D}_{i-1}$ and D_i . Note that data interaction can go both ways, either increasing or decreasing the probability of event $A = a$ occurring. Hence, ignoring data interaction by making $v_i = 1$ is not necessarily a conservative approximation.

Case $v_0 = 1$

The single v_0 parameter in (5) is noteworthy because

- it is data sequence-independent though involving jointly all n data events D_i ;
- the no-data interaction $v_0 = 1$ does not require that any or all elementary values v_i be equal to 1. In other words, multiple and possibly complex elementary data interactions ($v_i \neq 1$) may cancel out into a “global nondata interaction” $v_0 = 1$.

The approximation $v_0 = 1$ is different and far less restrictive than the joint conditional independence hypothesis written as

$$P(D_1, \dots, D_n | A = a) = P(D_1 | A = a) \cdots P(D_n | A = a).$$

Please see the case study in a subsequent section.

Dictatorial Property

If any of the elementary information D_i is fully decisive it imposes its result (as a dictator) to the final conditional probability, no matter the other data D_j , $j \neq i$ provided that none of them is contradictory and equally dictatorial.

Indeed, if $x_i = 0$ (∞) or equivalently if $P(A | D_i) = 1$ (0), then the final distance x given by the exact (5) is equal to 0 (∞). That is $P(A | D_i) = 1$ (0) leads to a decisive statement: $A = a$ occurs certainly (certainly not). A contradictory, but equally dictatorial data event D_j would yield $x_j = \infty$ when $x_i = 0$ and then (5) would not be able to resolve such total contradiction.

The limit case $v_i = 0$ corresponds to data interaction leading to decisive results. Indeed, first assume that none of the n data D_i is dictatorial by itself, i.e. $x_i \neq 0$ or ∞ , $\forall i = 1, \dots, n$. Then, $v_i = 0$ if and only if the numerator of its expression (4) is zero, that is

$$P(D_i = d_i | A = \text{nona}, \overline{D}_{i-1}) = 0, \quad \text{although} \quad P(D_i = d_i | A = \text{nona}) > 0.$$

In other words, the previous information $\overline{D}_{i-1}$ strongly indicates that the probability (likelihood) of observing the datum $D_i = d_i$ in presence of $A = \textit{nona}$ is zero, thus $v_i = 0$ entailing $v_i \frac{x_i}{x_0} = 0$, although $x_i > 0$, and finally

$$x = 0, \quad \text{that is} \quad P(A = a | D_i, i = 1, \dots, n) = 1.$$

Similarly, the case $v_i = \infty$ would entail $x = \infty$, hence $P(A = a | D_i, i = 1, \dots, n) = 0$, even though $x_i \neq \infty, \forall i$.

A Data Interaction Measure

Considering the more general notations $D_i, \Delta_j = \overline{D}_{i-1}$ and $v_{ij} = v_i$, the deviation $|1 - v_{ij}|$ is suggested as a measure of interaction between the two data sets D_i and Δ_j . In view of the symmetry of the two limit dictatorial cases $v_i = 0$ and $v_i = \infty$, a more symmetric and standardized measure of data interaction (when it comes to discriminating $A = a$ from $A = \textit{nona}$) could be

$$K_{ij} = 1 - \exp(-|\log v_{ij}|) = K_{ji} \in [0, 1], \quad (6)$$

where $v_{ij} = v_{ji} \geq 0$ is the nu parameter defined by (4) generalized by considering the notation $\Delta_j = \overline{D}_{i-1}$.

This interaction measure takes the limit values

- $K_{ij} = 1$, that is maximum data interaction, whenever $v_{ij} = 0$ or $+\infty$.
- $K_{ij} = 0$, that is no data interaction, whenever $v_{ij} = 1$ as discussed before.

Inference

Recall that knowing the distance x is equivalent to knowing the sought-after, fully conditioned probability through (1)

$$P(A = a | D_i = d_i, i = 1, \dots, n) = \frac{1}{1 + x} \in [0, 1].$$

The main contribution of this paper resides in the exact decomposition (5) of the fully conditioned distance x into

- The elementary distances: prior x_0 and single datum event-conditioned $x_i, i = 1, \dots, n$.
- The data interaction parameters $v_0, v_i, i = 1, \dots, n$.

This decomposition allows an inference paradigm whereby

1. Each of the elementary distances, x_0, x_i , or correspondingly, the elementary probabilities $P(A = a), P(A = a | D_i = d_i)$, would be evaluated directly from the d_i -data by whichever calibration technique is deemed appropriate for the study at hand.

2. The global interaction (correction) parameter $v_0(a; d_i = 1, \dots, n)$, a function of the a -value and the d_i -data values, could be lifted from some training data reflecting a proxy of the actual data interaction.

For example, one could generate approximative joint realizations of the $(n + 1)$ random events $\{A; D_i, i = 1, \dots, n\}$. Denote L such simulated realizations as $\{a^{(l)}; d_i^{(l)}, i = 1, \dots, n\}, l = 1, \dots, L$. From these realizations, one could retrieve a simulated and approximative correction parameter v_0^* function of a and of some summary statistics $\varphi_k(\cdot)$ of the n data values, say

$$v_0^*(a; \varphi_k(d_i, i = 1, \dots, n), k = 1, \dots, K), \quad (7)$$

where $K \ll n$ would provide a large dimension reduction.

In the case example given hereafter, $n = 3$ different data were summarized into their single average value ($K = 1$). The proxy function (7) would then be used to provide the approximate global interaction parameter v_0^* used to combine the elementary distances x_0, x_i , determined in step 1. This procedure can be extended to evaluating the n elementary interaction parameters $v_1^* = 1, v_i^*, i = 2, \dots, n$.

Note again that (7) is dependent on the outcome value $A = a$ being assessed. Were it changed from a to $a \neq \text{nona}$, then (7) should in general provide a different value v_0^* .

This parameter inference paradigm relies on the conjecture that the v_i -parameters being ratios of ratios (see (4)) would be reasonably stable, hence could be borrowed from proxy data sets, just like one would borrow a variogram/covariance function from a proxy outcrop/similar field data set. The elementary distances x_0, x_i , i.e., the elementary probabilities $P(A), P(A|D_i = d_i)$, however, would have to come from real data (not proxy), just like in kriging where data are real even though the variogram model may be borrowed.

The Tau Representation

The tau model introduced by Bordley (1982) and shown to be an exact representation of the fully conditioned probability by Krishnan (2005) is written

$$\frac{x}{x_0} = \prod_{i=1}^n \left(\frac{x_i}{x_0} \right)^{\tau_i} \quad \text{or} \quad \ln x - \ln x_0 = \sum_{i=1}^n \tau_i [\ln x_i - \ln x_0], \quad (8)$$

with

$$\tau_i = \frac{\log \frac{P(D_i|\tilde{A}, \bar{D}_{i-1})}{P(D_i|A, \bar{D}_{i-1})}}{\log \frac{P(D_i|\tilde{A})}{P(D_i|A)}} \in [-\infty, +\infty], \quad \tau_1 = 1.$$

The one-to-one relation with the nu representation (5) is obtained through the parameter relations

$$v_i = \left(\frac{x_i}{x_0} \right)^{\tau_{i-1}}, \quad \text{or} \quad \tau_i = 1 + \frac{\ln v_i}{\ln \frac{x_i}{x_0}}. \quad (9)$$

Recall that all these expressions are $(a, d_i; i = 1, \dots, n)$ values-dependent.

Tau or Nu Model?

If the v_i , τ_i parameters are evaluated explicitly, e.g., from training data, and made data-values dependent, (5) and (8) are equivalent. One should then prefer the nu-formulation as it puts forward a single correction parameter v_0 ($a, d_i, i = 1, \dots, n$) which is independent of the data sequence $D_1, D_2, \dots, D_n$. Also, evaluation of the τ_i parameter associated with noninformative datum such that $P(D_i|A) \cong P(D_i|\tilde{A})$ would run into the problem of division by a log ratio close to zero (see (8)).

However, if τ_i and v_i are assumed constant, independent of the data values ($a, d_i, i = 1, \dots, n$), then the tau formulation should be preferred. Indeed, consider the case of only two data events with the two different sets of data values

$$\{D_1 = d_1, D_2 = d_2\} \quad \text{and} \quad \{D_1 = d'_1, D_2 = d'_2\}.$$

- The nu model with the constant (homoscedastic) v_0 parameter value is written as

$$\begin{aligned} \frac{x}{x_0} &= v_0 \frac{x_1}{x_0} \frac{x_2}{x_0} \quad \text{for data set } \{d_1, d_2\}, \\ \frac{x'}{x_0} &= v_0 \frac{x'_1}{x_0} \frac{x'_2}{x_0} \quad \text{for data set } \{d'_1, d'_2\}, \end{aligned}$$

where x, x_1, x_2 are the distances corresponding to $\{d_1, d_2\}$ and x', x'_1, x'_2 are the distances corresponding to $\{d'_1, d'_2\}$. Conditional distances are data values-dependent, as opposed to the prior distance $x_0 = x'_0$. Therefore

$$\frac{x'}{x} = \frac{x'_1}{x_1} \frac{x'_2}{x_2}, \quad \text{for all constant } v_0 \text{ values.}$$

The parameter v_0 is seen to be ineffective.

- Conversely, the tau model with constant τ_1 and τ_2 parameter values is written as

$$\begin{aligned} \log \frac{x}{x_0} &= \tau_1 \log \frac{x_1}{x_0} + \tau_2 \log \frac{x_2}{x_0} \quad \text{for data set } \{d_1, d_2\}, \\ \log \frac{x'}{x_0} &= \tau_1 \log \frac{x'_1}{x_0} + \tau_2 \log \frac{x'_2}{x_0} \quad \text{for data set } \{d'_1, d'_2\}. \end{aligned}$$

Thus

$$\log \frac{x'}{x} = \tau_1 \log \frac{x'_1}{x_1} + \tau_2 \log \frac{x'_2}{x_2},$$

or equivalently

$$\frac{x'}{x} = \left(\frac{x'_1}{x_1} \right)^{\tau_1} + \left(\frac{x'_2}{x_2} \right)^{\tau_2}.$$

The tau parameters, although data values-independent, remain effective unless $\tau_1 = \tau_2 = \nu_0 = 1$. This latter property of the tau expression, remaining effective even if the τ_i 's are considered data values-independent, make the tau expression (8) a convenient heuristic to weight more certain data events. It suffices to make $\tau_i > \tau_j > 0$ to give more importance to data event D_i as compared to data event D_j , whatever the actual data value (d_i, d_j). This heuristic utilization of the tau model completely misses the main contribution of the tau/nu expression which is the quantification of data interaction for any specific set of values ($a, d_i; i = 1, \dots, n$).

An Example of Data Interaction

To investigate how the nu and tau parameters relate to data interaction, the following simple experiment is proposed. It involves one unknown A located at the center of an equilateral triangle with 3 data I_1, I_2, I_3 located at its three apexes (Fig. 1). The field is binary (0, 1) and has been generated by truncation of a Gaussian field.

More precisely, 100,000 unconditional joint realizations of four corresponding Gaussian random variables are generated by LU decomposition of their covariance matrix (program *LUSIM* in Deutsch and Journel 1998).

The isotropic covariance model is

$$\gamma(\mathbf{h}) = \exp\left(-\frac{|\mathbf{h}|}{r}\right), \quad \text{with practical range } 3r.$$

That range $3r$ is made variable from one set of 100,000 realizations to another. Each set allows the study of data dependence and interaction in evaluating the central value A .

Fig. 1 Spatial locations of three data I_1, I_2, I_3 and the unknown A . The distances are given in parentheses

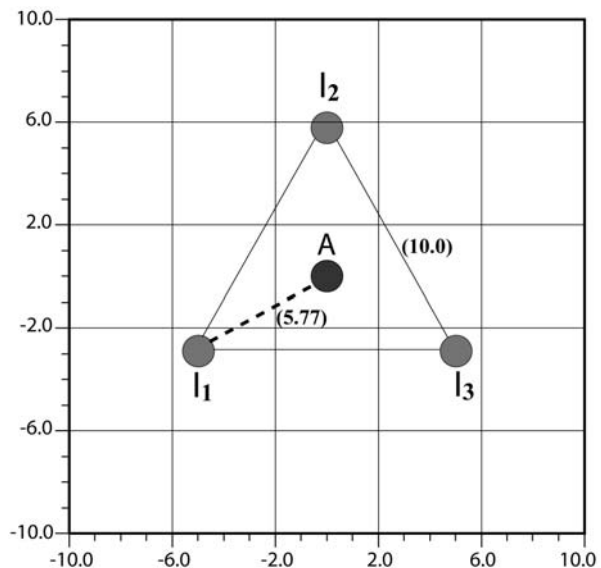


Table 1 Probabilities for the 16 joint occurrences

	<i>A</i>	<i>I</i> ₁	<i>I</i> ₂	<i>I</i> ₃		<i>A</i>	<i>I</i> ₁	<i>I</i> ₂	<i>I</i> ₃
<i>p</i> ₁	0	0	0	0	<i>p</i> ₉	1	0	0	0
<i>p</i> ₂	0	0	0	1	<i>p</i> ₁₀	1	0	0	1
<i>p</i> ₃	0	0	1	0	<i>p</i> ₁₁	1	0	1	0
<i>p</i> ₄	0	0	1	1	<i>p</i> ₁₂	1	0	1	1
<i>p</i> ₅	0	1	0	0	<i>p</i> ₁₃	1	1	0	0
<i>p</i> ₆	0	1	0	1	<i>p</i> ₁₄	1	1	0	1
<i>p</i> ₇	0	1	1	0	<i>p</i> ₁₅	1	1	1	0
<i>p</i> ₈	0	1	1	1	<i>p</i> ₁₆	1	1	1	1

These standard Gaussian realizations, denoted $z(\mathbf{u})$, are truncated at the median value $z = 0$ to generate joint realizations of the four binary indicator variables

$$i(\mathbf{u}_\alpha) = \begin{cases} 1 & \text{if } z(\mathbf{u}_\alpha) > 0, \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

where $\mathbf{u}_0$ is the location of the central value A to be evaluated, and $\mathbf{u}_\alpha$, $\alpha = 1, 2, 3$ are the three data locations.

The four variables being binary, there are a total of $2^4 = 16$ possible joint combinations of their values. A joint probability of occurrence p_k , $k = 1, \dots, 16$ is assigned to each of these joint combinations (Table 1). Note that $\sum_{k=1}^{16} p_k = 1$.

The prior or marginal probability associated to any of the four binary variables is $p_0 = 0.5$ corresponding to the prior distance

$$x_0 = \frac{1 - p_0}{p_0} = 1 = P(A = 1). \quad (11)$$

These 16 probabilities p_k are set equal to the corresponding 16 proportions of joint occurrence calculated from each set of 100,000 simulated realizations of the four variables A , I_1 , I_2 , I_3 . From such an exhaustive set of probabilities of joint occurrence, all conditional probabilities can be retrieved (Fig. 2). For example, the probability that $A = 1$ given that all three indicator data are 1 is retrieved as

$$P(A = 1 | I_1 = I_2 = I_3 = 1) = \frac{P(AI_1I_2I_3 = 1)}{P(I_1I_2I_3 = 1)} = \frac{p_{16}}{p_8 + p_{16}}. \quad (12)$$

The practical range $3r$ of the underlying Gaussian random function is increased along the abscissa axis of Fig. 2 from zero (pure nugget effect) to 30, a large range value three times the distance between any two of the three data I_j in Fig. 1. The ordinate axis gives various probabilities, all related to the event $A = 1$. All of these probabilities are seen to increase with the correlation range $3r$, as expected for the case of concordant data: $I_1 = I_2 = I_3 = 1$. The probabilities plotted along the ordinate axis of Fig. 2 are:

- The constant prior probability $P(A = 1) = 0.5$.

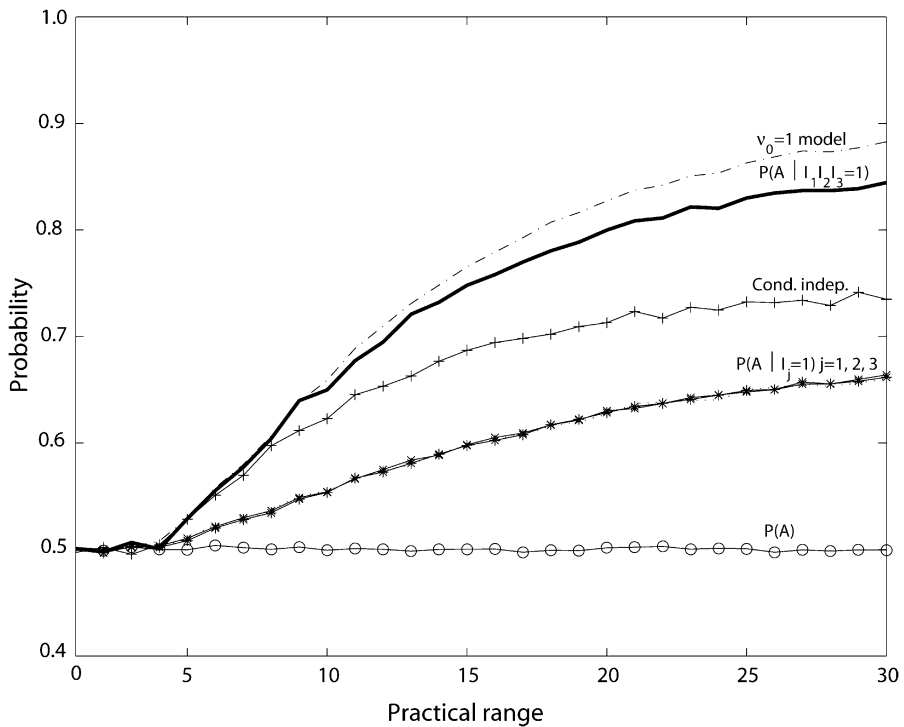


Fig. 2 Conditional probabilities. Concordant data case: $A = I_1 = I_2 = I_3 = 1$

- The three elementary conditional probabilities conditioned to one datum at a time: $P(A = 1|I_j = 1)$, $j = 1, 2, 3$. Had we drawn an infinite number of realizations (instead of only 100,000) these three elementary conditional probabilities would all be exactly equal.
- The exact fully conditioned probability $P(A = 1|I_1 I_2 I_3 = 1)$ as calculated from (12).
- The approximation of the previous exact fully conditioned probability using the nu expression (5) with the approximation $\nu_0 = 1$

$$P_{\nu_0=1}^*(A = 1|I_1 I_2 I_3 = 1) = \frac{1}{1 + x^*}, \quad (13)$$

with the distance x^* approximated with the $\nu_0 = 1$ model as

$$\frac{x^*}{x_0} = \prod_{k=1}^3 \frac{x_k}{x_0}, \quad x_k = \frac{1 - P(A = 1|I_k = 1)}{P(A = 1|I_k = 1)}.$$

- The approximation of the exact fully conditioned probability using the common hypothesis of data conditional independence. More precisely, the exact probability

is written

$$P(A = 1|I_1, I_2, I_3 = 1) = \frac{P(I_1 = I_2 = I_3 = 1|A = 1)P(A = 1)}{P(I_1 = I_2 = I_3 = 1)}$$

per data conditional independence

$$\begin{aligned} P(I_1 = I_2 = I_3 = 1|A = 1) \\ = P(I_1 = 1|A = 1)P(I_2 = 1|A = 1)P(I_3 = 1|A = 1) \end{aligned}$$

where, in terms of the probabilities p_k of Table 1,

$$P(I_1 = 1|A = 1) = \sum_{k=13}^{16} p_k \bigg/ \sum_{k=9}^{16} p_k,$$

$$P(I_1 = 1, I_2 = 1, I_3 = 1) = P(I_1 I_2 I_3 = 1) = p_8 + p_{16}.$$

Thus, the approximation provided by data conditional independence is

$$P_{CI}^*(A = 1|I_1 I_2 I_3 = 1) = \frac{P(A = 1) \prod_{k=1}^3 P(I_k = 1|A = 1)}{P(I_1 I_2 I_3 = 1)}. \quad (14)$$

Remark The denominator in (14), although the exact value is available here, is typically very difficult to get in real practice. This is the reason why the practice of conditional independence considers ratio of conditional probabilities of the type (Gelfand and Smith 1990)

$$\frac{P_{CI}^*(A = 0|I_1 I_2 I_3 = 1)}{P_{CI}^*(A = 1|I_1 I_2 I_3 = 1)} = \frac{P(A = 0)}{P(A = 1)} \prod_{k=1}^3 \frac{P(I_k = 1|A = 0)}{P(I_k = 1|A = 1)}. \quad (15)$$

This ratio is none other than (3) under data joint conditional independence given $A = 1$ and $A = 0$, i.e., (15) entails the $v_0 = 1$ model. However, the $v_0 = 1$ model does not call necessarily for the previous two conditional independence (given $A = 1$ and given $A = 0$); in that regard it is a less restrictive hypothesis.

Comments

From Fig. 2, the $v_0 = 1$ model is seen to provide a better approximation than the conditional independence hypothesis (given $A = 1$), increasingly better as the correlation range $3r$ increases. The $v_0 = 1$ model corresponds to a hypothesis of no data interaction which is increasingly poorer as the correlation within data, and between data and unknown increases. In the case of concordant data ($I_1 = I_2 = I_3 = 1$ used to evaluate the probability of $A = 1$), ignoring data interaction by assuming $v_0 = 1$ leads to overcompounding the 3 individual probabilities $P(A = 1|I_k = 1)$ leading to an overestimation increasing with correlation range.

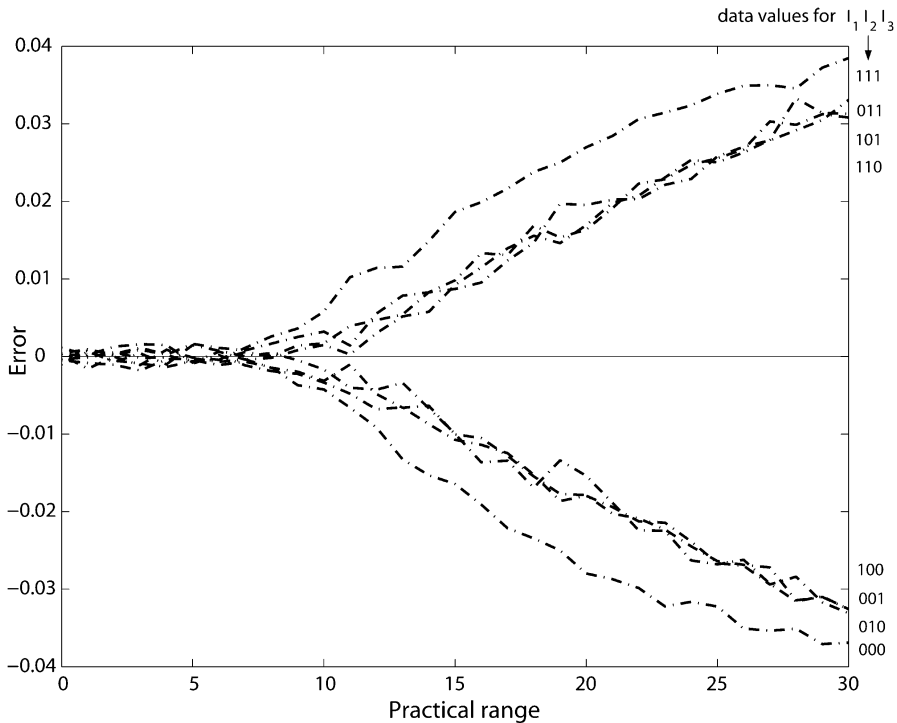


Fig. 3 Data values-dependent error associated to the $\nu_0 = 1$ model

Data Values Dependence

To evaluate how the $\nu_0 = 1$ approximation fares depending on the set of 3 data values, Fig. 3 plots the error

$$[P_{\nu_0=1}^*(A = 1|i_1, i_2, i_3) - P(A = 1|i_1, i_2, i_3)]$$

for the $2^3 = 8$ possible sets of data values ($I_1 = i_1, I_2 = i_2, I_3 = i_3$).

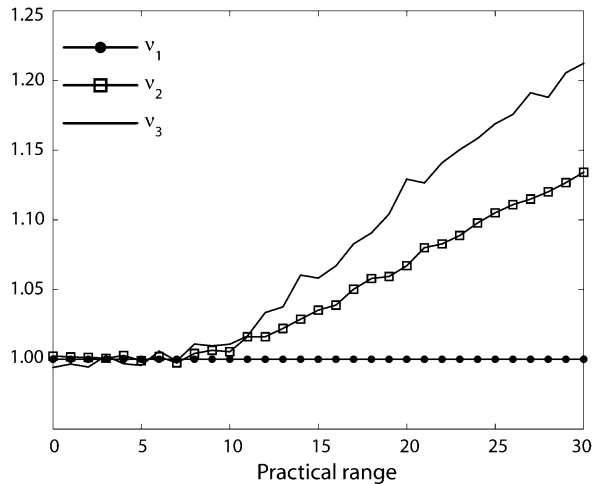
As expected, ignoring data interaction leads to increasing errors as the correlation range increases with overestimation when 2 or more of these three data are valued 1 and underestimation for the other cases.

Exact Nu Weights

Availability of the exhaustive set of 16 joint probabilities p_k allows calculation of the exact (ν_i, ν_0) weights as given in expression (4).

Figure 4 shows the three data sequence-dependent ν_i weights calculated for the case $A = I_1 I_2 I_3 = 1$. Since there are three data I_1, I_2, I_3 , there are $3! = 6$ possible data sequences. However, because of the equilateral data configuration (Fig. 1) associated with an isotropic correlation, the data sequence does not matter here. The first

Fig. 4 The sequence-dependent v_i weights for the data concordant case $A = I_1 = I_2 = I_3 = 1$



datum in any sequence always receives a unit weight $v_1 = 1$. As the correlation range increases, the interaction between the two first data increases leading to an increasing second nu weight v_2 . The third datum v_3 reflects the even greater interaction between the first two data and the last one. Recall that data interaction, hence the nu weights, is data values-dependent; that interaction is maximal here when all data are concordant $I_1 = I_2 = I_3 = 1$, and it increases with the correlation range.

Figure 5 gives the single, data sequence-independent, exact v_0 weight. The case considered here is again that of concordant data $I_1 = I_2 = I_3 = 1$. The exact v_0 weight is seen to be close to 1 when the practical range is smaller than the datum-to-unknown distance 5.77, and it increases thereafter with the correlation range.

Our inference paradigm, as presented before, consists of two steps

1. Evaluate the single datum-conditioned probabilities $P(A = 1|I_j = 1)$, $j = 1, 2, 3$ using calibration data from the actual field under study.
2. Borrow the data values-dependent v_0 weight from some training data or expert catalog and export it to the actual field under study to combine the previous single datum-conditioned probabilities.

For example, assume that from some prior expertise (perhaps built from experiments on training data sets similar to those used in this study), we have access to the following v_0 weight function

$$v_0 = \begin{cases} 1 & \text{whatever the data values for all range } 3r < 6, \\ 1 + \frac{(3r-6)^2}{2000} & \text{as function of the practical range } 3r > 6. \end{cases}$$

This function is assumed applicable only when two or more data are valued 1. The error graphs of Fig. 3 are recalculated using this improved v_0 -model. The results of Fig. 6 show a significant decrease in the error and demonstrate our point that the worth and practicality of the nu/tau parameters depend on our ability to go beyond the approximation $v_0 = 1$. This calls for calibrating the v_0 -values for classes of data values.

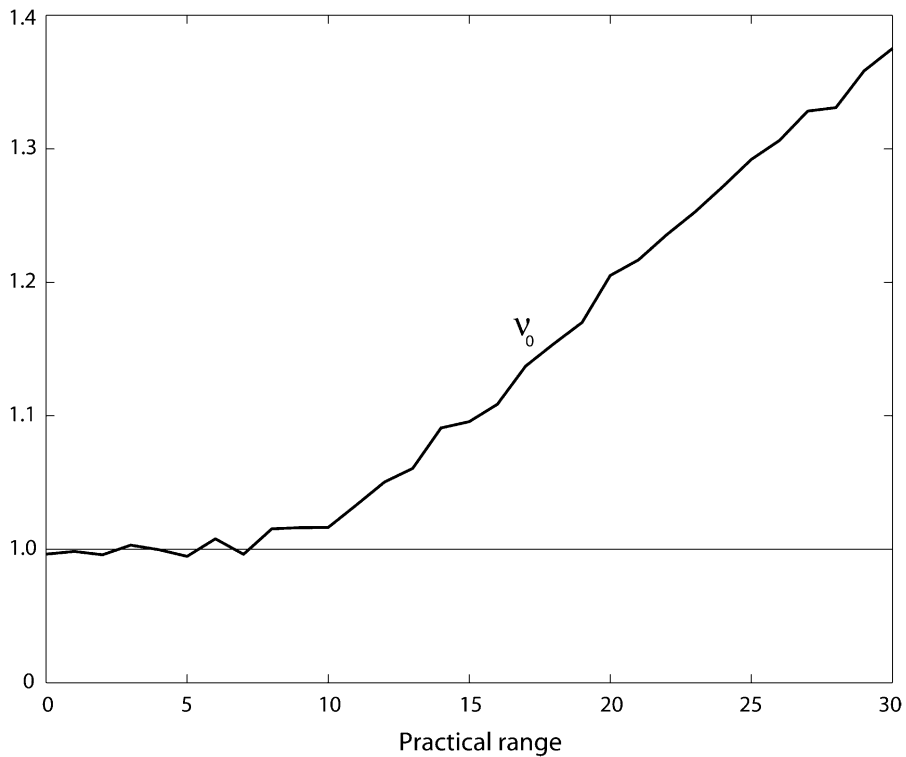


Fig. 5 The single sequence-independent v_0 weight for the data concordant case $A = I_1 = I_2 = I_3 = 1$ for sequence $I_1 \rightarrow I_2 \rightarrow I_3$

Conclusions

This study allows us to draw several preliminary conclusions which may have a significant impact on how we address the problem of data integration in the future. Combining the contributions of data events from different sources is not a simple task. Each of these data events could be quite complex and their contributions are generally related to one another, thus any data event can influence the information brought by other data or combination of data.

The remarkably simple $v_0 = 1$ model verifies all limit properties of probabilities and has been shown to consistently outperform traditional models based on some form of data independence. The contribution of the nu representation compared to the tau representation is to express the impact of joint data interaction by a single correction parameter v_0 whose known exact expression can lead to case-dependent approximations. We suggest an inference paradigm where v_0 -values would be borrowed from a catalog of “similar” cases where the v_0 -values are determined from training data for classes of “similar” data configurations and data values. The future practicality of the nu model would depend on our ability to generate such proxy training data from which to borrow and export the v_0 -parameters.

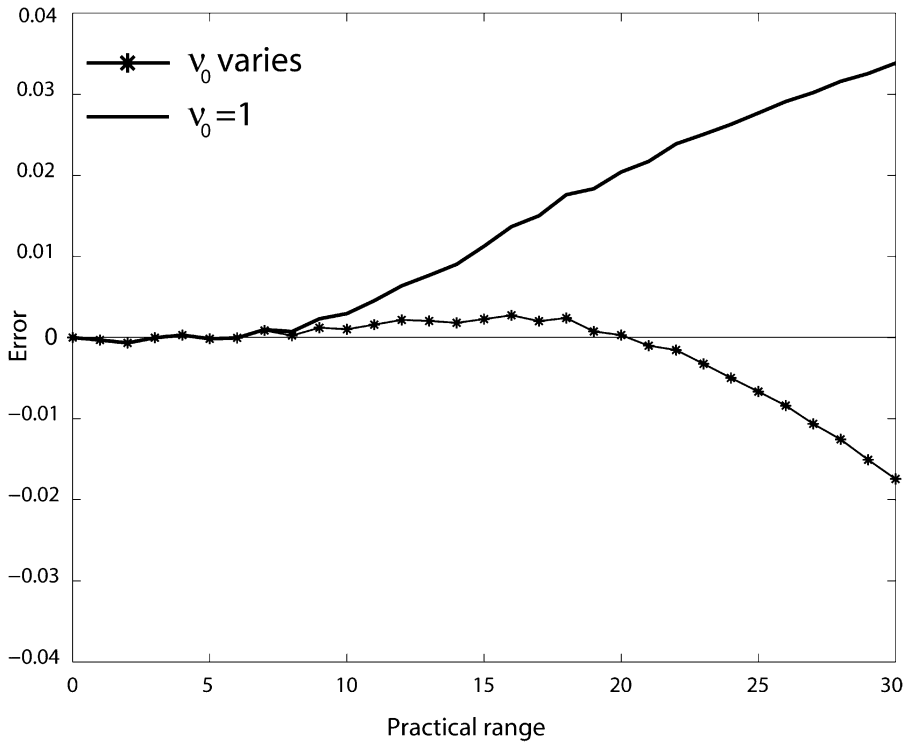


Fig. 6 The averaged error associated with a data value-dependent v_0 model and with the $v_0 = 1$ model

The theoretical developments presented in the first part of this paper have led to an attempt at defining a measure of data interaction given a particular estimation goal. Data interaction is shown to be both data values and goal-dependent. Joint no data-interaction, i.e., the $v_0 = 1$ model, is a concept much richer and more useful than that of data conditional independence.

Appendix: An Analytical Example

This small analytical example demonstrates the difference between the concept of data conditional independence and data interaction.

Consider the following general function of two binary data I_1 and I_2

$$A = \lambda_0 + \lambda_1 I_1 + \lambda_2 I_2 + \lambda_3 I_1 I_2$$

λ_0 is but a translation parameter and can be set to zero.

Table 2 Outcomes of a function of 2 binary data and their probabilities

	$I_1 I_2 = 1$	$I_1(1 - I_2) = 1$	$(1 - I_1)I_2 = 1$	$(1 - I_1)(1 - I_2)$
$A =$	$\lambda_1 + \lambda_2 + \lambda_3$	λ_1	λ_2	0
with probability	p_{12}	$p_1 - p_{12}$	$p_2 - p_{12}$	$1 - p_1 - p_2 + p_{12}$

The joint distribution of the two binary (0, 1) data I_1 and I_2 is fully characterized by the three probabilities

$$p_1 = E\{I_1\} = P(I_1 = 1),$$

$$p_2 = E\{I_2\} = P(I_2 = 1),$$

$$p_{12} = E\{I_1 I_2\} = P(I_1 = 1, I_2 = 1).$$

The two data I_1 and I_2 are independent if $p_{12} = p_1 p_2$.

The variable A can take only 4 values depending on the joint outcome (i_1, i_2) of the two binary data (Table 2). The outcome $A = \lambda_1$, entails necessarily $I_1(1 - I_2) = 1$, hence

$$\begin{aligned} P(I_1 = 1|A = \lambda_1) &= P(1 - I_2 = 1|A = \lambda_1) \\ &= P(I_1(1 - I_2) = 1|A = \lambda_1) = 1 \quad \forall p_{12}. \end{aligned}$$

Thus the two data I_1 and I_2 are conditionally independent given $A = \lambda_1$, whether or not they are independent.

Now, consider the complementary event $A \neq \lambda_1$. From Table 2, it comes

$$\begin{aligned} P(I_1 = 1|A \neq \lambda_1) &= \frac{p_{12}}{1 - (p_1 - p_{12})} = \frac{p_{12}}{1 + p_{12} - p_1}, \\ P(I_2 = 1|A \neq \lambda_1) &= \frac{p_{12} + (p_2 - p_{12})}{1 - (p_1 - p_{12})} = \frac{p_2}{1 + p_{12} - p_1}, \\ P(I_1 I_2 = 1|A \neq \lambda_1) &= \frac{p_{12}}{1 + p_{12} - p_1} = P(I_1 = 1|A \neq \lambda_1), \quad \forall p_{12}. \end{aligned}$$

Excluding the trivial case $p_2 = 1$, there is no conditional independence of I_1, I_2 given $A \neq \lambda_1$, no matter what the value of p_{12} is.

This simple example demonstrates that

- Conditional independence and data independence (or dependence) are unrelated. This is trivial; the former includes the trivariate distribution (A, I_1, I_2) , the latter only its marginal bivariate distribution (I_1, I_2) . One can have data dependence (or independence) and simultaneously conditional independence.
- Data independence and, simultaneously, conditional independence given $A = a$ does not necessarily entail conditional independence given $A = \text{nona}$.

References

- Benediktsson, Swain (1992) Consensus theoretic classification methods. *IEEE Trans Systems Man Cybern* 22:688–704
- Bordley RF (1982) A multiplicative formula for aggregating probability assessments. *Manag Sci* 28(10):1137–1148
- Caers J, Hoffman TB (2006) The probability perturbation method: an alternative Bayesian approach for solving inverse problems. *Math Geol* 38(1):81–100
- Deutsch C, Journel AG (1998) *GSLIB: Geostatistical software library and user's guide*. Oxford University Press, New York
- Fleiss JL (1981) *Statistical methods for rated and proportions*, 2nd edn. Wiley, New York
- Galton F (1894) *Natural Inheritance*, 5th edn. Macmillan, New York
- Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc* 85:398–409
- Genest C, Zidek JV (1986) Combining probability distributions: A critique and an annotated bibliography. *Stat Sci* 1:114–118
- Journel AG (1983) Nonparametric estimation of spatial distributions. *Math Geol* 15(3):793–806
- Journel AG (2002) Combining knowledge from diverse data sources: an alternative to traditional data independence hypothesis. *Math Geol* 34(5):573–596
- Krishnan S (2005). Combining diverse and partially redundant information in the earth sciences. PhD thesis, Department of Geological and Environmental Sciences, Stanford University, USA
- Strebel S (2001) Conditional simulation of complex geological structures using multiple-point statistics. *Math Geol* 34(1):1–22