Combining probabilities

prepared by: Coryn A.L. Bailer-Jones, Kester Smith
Max Planck Institute for Astronomy, Heidelberg
Email: calj@mpia.de
approved by:
reference: GAIA-C8-TN-MPIA-CBJ-053
issue: 2
revision: 1
date: 2011-12-20
status: Issued

Abstract

We show how to combine posterior probabilities from an ensemble of models, each of which estimates the same parameter (or class) but using "independent" data. From this we describe how to separate out and replace the class prior (or the model-based prior) of a classifier post hoc and show how this relates to the combination problem. We also discuss the subtleties of conditional independence, what "independent data" means, and outline under what circumstances dependent variables can become independent when conditioned on new information.

Contents

Document History

Issue	Revision	Date	Author	Comment
2	1	2011-12-20	CBJ	Typo corrected (thanks to Thomas Mülders for spotting this)
2	0	2011-07-19	CBJ	Comments from David Hogg, Chao Liu and KS. Sections 1.0 and 1.2 modified. Other minor revisions. Issued.
2	D	2011-07-14	CBJ	Sections 2.1 and 3.2 corrected: unconditional independence removed resulting in a class-independent normlization constant. Sections 4.2 and 4.3 re-written. Other minor revisions.
1	0	2010-01-21	CBJ	Minor changes. Issued.
D	1	2010-01-15	CBJ	Draft based on discussions with KS

1 Basics

The goal of probabilistic inference is to estimate the probability density function of some parameter based on observed data. Certainty is not guaranteed because the data are noisy. Sometimes we may make multiple estimates of the same parameter using different, perhaps "independent", measurements. We examine here the problem of how to combine these under various conditions. In the interests of being didactic we develop the ideas in some detail from first principles.

It should be emphasized from the very beginning that the concept of "independence" is a tricky one, and simply stating that two things are "independent" without any qualification is an almost meaningless statement. It is one of the goals of this TN to clarify what we mean by independence, or rather, to emphasize that if we say something is "independent" then we should always qualify this (mathematically).

1.1 Bayes' theorem

A fundamental and familiar theorem is Bayes' theorem, which relates two conditional probabilities of quantities A and B

$$P(A|B) \;=\; \frac{P(B|A)\;P(A)}{P(B)} \tag{1}$$

which follows from $P(A,B) = P(A|B)P(B) = P(B|A)P(A)$. If we consider A as the parameter of interest, and B as the data, then we often refer to $P(A|B)$ as the *posterior probability* of parameter A given a measurement of data B, $P(B|A)$ as the *likelihood* of getting data B given parameter A, and $P(A)$ as the *prior probability* of A. For example, A could be the $T_{\rm eff}$ of a star and B the observed spectrum. Alternatively, A could be a discrete class, such as "quasar". A model for $P(A)$ can involve any information not included in B, which is useful for estimating

A. For the quasar example this might be the apparent magnitude (there are very few bright quasars).

Note that $P(B) = \int_A P(B|A)P(A)dA$ is the prior probability of observing data B. We are not interested in this quantity in the current context, i.e. once we have already measured the data (but it is important in the context of model comparison, where it is called the *evidence*).

1.2 What do classifiers provide?

Consider a two-way classification problem in which we want to estimate the probability of class C based on data D. $\overline{C}$ is the complement of class C. A classifier such as a support vector machine (when set up to provide probabilities) provides the posterior probability $P(C|D)$, rather than the likelihood of getting the data given the class, $P(D|C)$. This must of course be the case if we assign one minus this output as the probability of the other class, i.e. $P(\overline{C}|D) = 1 - P(C|D)$. In contrast $P(D|C) + P(D|\overline{C}) \neq 1$ in general.[1]

Some classification models are constructed to provide likelihoods. For example, the kernel density method models the density of the data for each class separately, so for a two-class problem it is clearly providing $P(D|C)$ and $P(D|\overline{C})$. If we only use these two quantities in the inference then we are implicitly using equal class priors, $P(C) = P(\overline{C}) = 1/2$ (assuming the density estimates are separately normalized for each class). In such cases we use the ratio of these two likelihoods, which is related to the posterior via Bayes' theorem

$$\begin{aligned} P(C|D) &= \frac{P(D|C)\,P(C)}{P(D)} \\ &= \frac{P(D|C)\,P(C)}{P(D|C)\,P(C) + P(D|\overline{C})\,P(\overline{C})} \\ &= \frac{1}{1 + \frac{P(D|\overline{C})\,P(\overline{C})}{P(D|C)\,P(C)}} \quad . \end{aligned} \tag{2}$$

The priors cancel if equal.

1.3 No important distinction between priors and posteriors

There is nothing special about a prior. For example, we could write Bayes' theorem[2] involving three parameters by conditioning everything on a third parameter (or measurement, or assump-

[1] Some people confuse $P(D|C)$ and $P(C|D)$ at first. A simple example will help. Imagine a country in which all cars are red, but half of all trucks are red and the other half are blue. You observe a red vehicle on the road but don't see whether it's a car or truck (there are no other vehicles). The observation, D, is "red", and let C be "car". $P(D|C) = 1$ (all cars red) and $P(D|\overline{C}) = 1/2$, which clearly don't add to unity. In contrast, the vehicle must be either a car or a truck, so $P(C|D) + P(\overline{C}|D) = 1$ whatever D is.

[2] You will of course get the same result if you instead replace B with (B, E) in equation 1. We can always introduce a new conditioning parameter in this way, provided we introduce it to every term.

tion etc.) E,

$$P(A|B,E) \;=\; \frac{P(B|A,E)\,P(A|E)}{P(B|E)} \quad . \tag{3}$$

In the quasar example B could be the apparent magnitude and E could be some additional information, such as a particular model for the universe. In that case we might think of $P(B|E)$ as a magnitude-based classifier, which obviously depends on the universe model (quasar luminosity function and distances). In that case we would think of $P(E)$ as being the prior probability of this particular model.

The point is that equation 3 is a general equation for *updating probabilities*. If we initially have an estimate of the probability of A based only on information E, i.e. $P(A|E)$, and then we measure new information B, Bayes' theorem gives us a way of updating our inference to give a probability of A based on both B and E, $P(A|B,E)$. The posterior from the first inference, $P(A|E)$, becomes the prior of the next. It is therefore often more useful to think in terms of "independent pieces of information" expressed as conditional probabilities, rather than talking of "priors" and "posteriors".

However, as it stands equation 3 is not terribly useful for this updating, so let's look at this in another way.

2 Updating probabilities (combining classifiers)

2.1 Method for independent classifiers

Consider that we have two sets of information, D_A and D_B, and we use each to independently (separately) assess some parameter (or class probability) C. For example D_A and D_B might be different spectra, or D_A could be a normalized spectrum and D_B the source apparent magnitude or the astrometry. We have (classification) models which estimate $P(C|D_A)$ and $P(C|D_B)$. How do we combine them to give $P(C|D_A,D_B)$, our best estimate based on both pieces of information?

Using Bayes' theorem

$$P(C|D_A,D_B) \;=\; \frac{P(D_A,D_B|C)\,P(C)}{P(D_A,D_B)} \tag{4}$$

If D_A and D_B are "independent measurements", then typically they are (or rather, we mean

they are) *conditionally independent* given the class, i.e.[3]

$$P(D_A, D_B|C) \;=\; P(D_A|C)\,P(D_B|C) \tag{6}$$

so equation 4 can be written

$$P(C|D_A, D_B) \;=\; \frac{P(D_A|C)\,P(D_B|C)\,P(C)}{P(D_A, D_B)} \quad . \tag{7}$$

Using Bayes' theorem to rewrite $P(D_A|C)$ as $P(C|D_A)P(D_A)/P(C)$ and similarly for $P(D_B|C)$, we get

$$\begin{aligned} P(C|D_A, D_B) \;&=\; \frac{P(D_A)P(D_B)}{P(D_A, D_B)} \times \frac{P(C|D_A)P(C|D_B)}{P(C)} \\ &=\; a\,\frac{P(C|D_A)P(C|D_B)}{P(C)} \end{aligned} \tag{8}$$

which defines a: It is a class-independent term depending only on the data. As we've measured these data, its value is not interesting here (we are not doing model comparison), so we treat it as a normalization constant which ensures that $\sum_k P(C_k|D_A, D_B) = 1$. Equation 8 is our final result for combining results from two independent classifiers (where "independent" here refers to the data being independent in the sense of equation 4). We can easily extend this to include a third independent piece of information, D_E,

$$\begin{aligned} P(C|D_A, D_B, D_E) \;&=\; P(D_A, D_B, D_E|C)\,\frac{P(C)}{P(D_A, D_B, D_E)} \\ &=\; P(D_A|C)\,P(D_B|C)\,P(D_E|C)\,\frac{P(C)}{P(D_A, D_B, D_E)} \\ &=\; \frac{P(C|D_A)P(D_A)}{P(C)}\,\frac{P(C|D_B)P(D_B)}{P(C)}\,\frac{P(C|D_E)P(D_E)}{P(C)}\,\frac{P(C)}{P(D_A, D_B, D_E)} \\ &=\; a\,\frac{P(C|D_A)\,P(C|D_B)\,P(C|D_E)}{P(C)^2} \end{aligned} \tag{9}$$

where a is a new normalization constant. In general, if we have N independent classifiers $n = 1 \ldots N$ each using independent information D_n, then they can be combined as

$$P(C|D_1, \ldots, D_N) \;=\; a\,\frac{\prod_{n=1}^{n=N} P(C|D_n)}{P(C)^{N-1}} \quad . \tag{10}$$

This is the equation to use when combining multiple independent classifiers.[4] If the D_n are not independent (conditioned on C), then in general we have to know their joint probability

[3] In contrast, they are not *unconditionally independent* in general, i.e.

$$P(D_A, D_B) \;\neq\; P(D_A)\,P(D_B)\,. \tag{5}$$

Note also that D_A and D_B can still be conditionally independent (equation 4) if they are different measurements of the same thing. Both of these points are discussed in section 4.

[4] The unconditional independence of D_A and D_B was incorrectly also assumed in version 1 of this document, which results in the normalization constant in equation 10 being omitted (i.e. $a = 1$, as the data priors then cancel). A simple numerical example shows that $\sum_k P(C_k|D_A, D_B) \neq 1$ in general if we assume unconditional independence.

distribution. Note that all of the classifiers, $P(C|D_n)$, as well as the prior $P(C)$, are implicitly assumed to be conditioned on the same background information.

2.2 Class fraction prior

There may arise some confusion as to the role of the "class fraction" prior when combining models. The class fraction for a given class is the overall fraction of objects of that class in a population. It may be known for a specific population (e.g. if 70 objects of a sample of 100 are stars, then the star class fraction is 0.7), or, more likely, it may be postulated for a general population based on more general information (e.g. the fraction of all stars in the universe brighter than G=19). The class fraction is independent of any specific data we otherwise use in the inference.

Take the case of classifying stars and quasars based on two classification models – one based on spectra, D_A, and another based on magnitude, D_B – and we want to introduce the overall class fraction of quasars to stars (independent of the magnitude or spectrum), which we take to be 1/100. How do we introduce this class fraction "prior", which we will call information D_E? We can simply consider this "prior" as a classifier which delivers $P(C|D_E)$, albeit it an extremely simple one because D_E is the very simple piece of data "quasars are 100 times rarer than stars". So if C is the class "star" then $P(C|D_E) = 100/101$. We can then use equation 9. In that case, $P(C)$ is the "prior" probability of the object being class C independent of D_A, D_B or D_E. This probability needs to be equal to what the models for $P(C|D_A)$, $P(C|D_B)$ and $P(C|D_E)$ have assumed (perhaps implicitly) this prior to be. That is, if the classifier based on the spectrum has an implicit prior for class C of $Q(C)$, then we require that $Q(C) = P(C)$, and likewise for the other two classifiers using D_B and D_E. We can see this more easily if we write down equation 9 explicitly conditioned on the background information, H

$$P(C|D_A, D_B, D_E, H) \;=\; a\,\frac{P(C|D_A, H)\,P(C|D_B, H)\,P(C|D_E, H)}{P(C|H)^2}\;. \qquad (11)$$

We now see that all models have to be conditioned on the same background information, H, as is the original prior, $P(C|H)$. It is often the case (but not necessarily so), that when we express the class fraction explicitly via D_E, then H just represents "no additional information", in which case $P(C|H)$ is a uniform prior. In the case of classification with K classes ($k = 1 \ldots K$), this means $P(C_k|H) = 1/K$. This *may* be appropriate for classifiers trained on balanced data sets (equal class fractions), but is not necessarily so. See Bailer-Jones et al. (2008) section 2.3.2 for a discussion.

Alternatively, we may have the class fraction "prior" already built into the background information H and therefore into the classifiers for D_A and D_B. Thus the values $P(C|D_A, H)$ and $P(C|D_B, H)$ will differ from $P(C|D_A)$ and $P(C|D_B)$ by some constant multiple. In that case we don't need an explicit model for $P(C|D_E, H)$, and $P(C|H)$ now includes our "class fraction prior". In equation 11 D_E is the same information as H, so $P(C|D_E, H) = P(C|H)$ and equation 11 reduces to equation 8 with all terms conditioned on H.

We have used this method in Smith et al. (2010) to identify Blue Horizontal Branch stars in

SDSS data.

3 Replacing prior information

3.1 General method

Sometimes a classification model infers the probability of a class C (or distribution over a parameter C) based on both some explicit data D_A and on a prior (assumptions, information etc.) D_o, yet we would like to modify this prior post hoc. Specifically, we may want to *replace* the (often implicit) prior of the model with a different explicit prior. This may be desirable if we train a classifier on balanced data (equal class fractions) but then want to apply it to a population which we know has very unbalanced class fractions, but without having to retrain the model. This was discussed in detail by Bailer-Jones et al. (2008) who introduced a very simple method for replacing the prior, which we summarize here (see sections 2.3 and 2.4 of that paper).

As the classification output (posterior) is proportional to the product of a likelihood and a prior, then if we know (or can work out) this "old" prior, then we can simply divide by this and multiply by the new prior, to achieve a modified model. Let $P(C|D_o)$ be the old prior, $P(C|D_A, D_o)$ the classification outputs from the model which is based on D_o, and $P(C|D_n)$ the new prior. The posterior probability for the modified model is

$$P(C|D_A, D_n) \;=\; a\, P(C|D_A, D_o)\, \frac{P(C|D_n)}{P(C|D_o)} \tag{12}$$

where a is a normalization constant to ensure that $\sum_k P(C_k|D_A, D_n) = 1$ (see appendix A; equation 12 is for any one of these classes C_k). This constant is required because in general the two priors for a given class will not follow the same normalizaton. In Bailer-Jones et al. (2008) we used the class fractions of the populations which D_o and D_n represent as proxies for the priors (and we showed how $P(C|D_o)$, the model-based prior, can be calculated from the outputs of the original model). D_o and D_n may be quite distinct, e.g. equal and non-equal class fractions respectively. Alternatively, D_o may represent a (constant) non-equal class fraction, and D_n could reflect a class fraction which depends on magnitude. If, when averaged over all magnitudes, D_n gave the same non-equal class fraction as given by D_o, then D_n *includes* D_o. Equation 12 holds whether or not D_n includes D_o (but the value of a may vary).

3.2 Equivalence of prior combination with prior replacement when the new prior includes the old

Equation 12 – in which we *replace* D_o with D_n – can be compared with equation 8 – in which we *combine* D_A with D_B. Let us re-write the latter equation swapping the symbol D_B with D_n and conditioning the whole thing on D_o

$$P(C|D_A, D_n, D_o) \;=\; a\, \frac{P(C|D_A, D_o)\, P(C|D_n, D_o)}{P(C|D_o)} \quad . \tag{13}$$

This holds whether or not D_n and D_o are unconditionally independent. Let us now assume that D_n includes D_o, i.e. $\{D_n, D_o\} \rightarrow D_{n'}$ (D_o does not include D_n). The equation becomes

$$P(C|D_A, D_{n'}) \;=\; a\,\frac{P(C|D_A, D_o)\,P(C|D_{n'})}{P(C|D_o)} \quad . \tag{14}$$

This is the same as equation 12. Thus if the "new" prior includes the "old" prior, then combining the model based on the old prior with the new prior is equivalent to replacing the priors.

4 Conditional independence

When are measurements "independent"? Why did conditional independence (equation 4) apply in the classifier combination, but not unconditional independence?

4.1 Repeated measurements

We first look at repeated measurements. Suppose we make two noisy measurements, D_1 and D_2, of the same quantity C. These might be two measurements of the flux of a star, for example. Consider two cases.

A. If we have no knowledge of and make no model of the value of C, then D_1 and D_2 are not independent, because we know they are of the same thing; so D_1 is our best estimate of D_2 (and vice versa).

B. If, instead, we assume some value for – or model of – C, then we know that D_1 and D_2 differ from C only by the noise. If the noise in the two measurements is independent, then these two measurements are independent conditional on C.

In case A our measurement D_2 tells us something about C and therefore about D_1, because actually D_1 and D_2 depend on C by construction. But in case B we condition on C and so already account for what's common between the two measurements, so they are conditionally independent. Put mathematically, B says $P(D_1, D_2|C) = P(D_1|C)P(D_2|C)$. In contrast A says $P(D_1, D_2) \neq P(D_1)P(D_2)$, i.e. they are not (unconditionally) independent.[5]

The normal situation is that we assume some model for our measurement process, e.g. a Gaussian with mean C (unbiased measurement) and variance V, and let us assume that V is known (the measurement precision). In that case our model for an observation D_i is the Gaussian probability distribution $P(D_i|C, V)$. As this is conditioned on C (and V), then N such measurements, $\{D_N\} = (D_1, \ldots, D_N)$, are independent and so can be combined to estimate the

[5]One might argue that case A does not represent repeated measurements, because if D_1 and D_2 are repeated measurements then they must be conditioned on this assumption.

parameter using Bayes' theorem

$$\begin{aligned} P(C|\{D_N\},V) &= \frac{P(C|V)\ P(\{D_N\}|C,V)}{P(\{D_N\}|V)} \\ &= \frac{P(C)\ \prod_{i=1}^{i=N} P(D_i|C,V)}{P(\{D_N\})} \ . \end{aligned} \tag{15}$$

4.2 General case

In general D_1 and D_2 may be conditionally independent (given C), even if they are not repeated measurements, as a simple example will show.[6]

Two students take different exams on Monday morning. Let D_1 and D_2 be the events that students 1 and 2 respectively get an A grade. We might expect these to be independent in the sense of $P(D_1, D_2) = P(D_1)P(D_2)$. However, it turns out that both students are rugby fans and that there was a late night rugby match on Sunday night, and attending the match (event C) reduces the probability of getting an A grade (which means $P(D_i|C) < P(D_i|\overline{C})$). If we ignored the information C, and then found out that one student did poorly in her exam, this would increase the probability (our degree of belief) that there had been a rugby match, which in turn increases the probability that the other student also did poorly. That is, D_1 and D_2 become dependent on each other. (We could imagine repeating this many times and finding a correlation between the scores of the two students.) In other words, the "hidden" (or ignored) information makes D_1 and D_2 mutually dependent, i.e. $P(D_1, D_2) \neq P(D_1)P(D_2)$ in general.

In contrast, given (taking into account) that C occurs, it is reasonable to assume that the performance of the two students is again independent (the rubgy diversion affects both students equally). That is, the students' performances are independent *conditional on* C. The reason is that if we already know C, then also knowing D_2 doesn't provide any *additional* information about the probability of D_1. (Conversely, if we leave out the information C, we can't be sure that D_1 and D_2 are independent: there may be a hidden variable connecting them.) Putting this mathematically: $P(D_1, D_2|C) = P(D_1|D_2, C)P(D_2|C)$ in general. If D_2 doesn't provide any more information about D_1 beyond what C provides, then $P(D_1|D_2, C) = P(D_1|C)$, and so $P(D_1, D_2|C) = P(D_1|C)P(D_2|C)$, the case of conditional independence. This is the usual case in scientifc inference, where we set up a model for the data, and look at the probabilities of the data conditioned on that model (or parameter of the model). ($P(D|C)$ may be a noise model, for example.) This applies to the discussion in section 2.

However, whether or not conditional independence applies depends subtly what we are conditioning on. Let D_1 be the measurement of the parallax of an object and D_2 the measurement of its apparent magnitude. These statements are conditional on this being the same object, the "hypothesis" H_0. Assuming both quantities to be unconstrained and independent, then $P(D_1, D_2|H_0) = P(D_1|H_0)P(D_2|H_0)$.

[6] This example has been adapted from http://cnx.org/content/m23258/latest/, where there are also numerical examples which illustrate these points further.

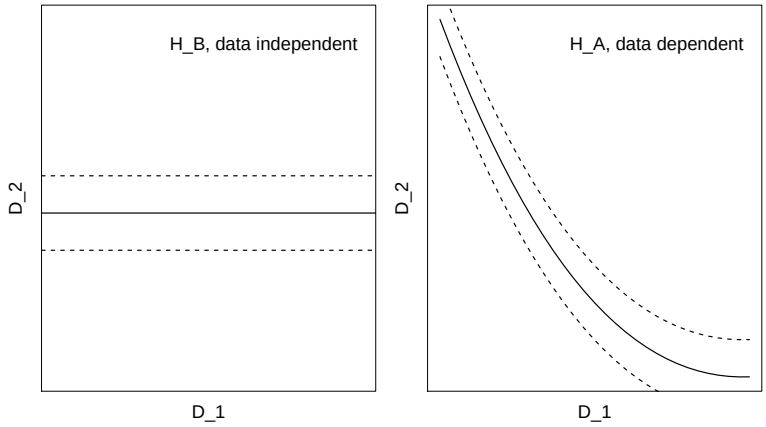

FIGURE 1: *Left:* Hypothesis H_B which limits D_2 but in which D_1 and D_2 are independent. *Right:* Hypothesis H_A in which D_1 and D_2 are mutually dependent. The solid line represents the expectation of the relationship and the dashed lines the 95% confidence interval (noise model) for the data.

Suppose we now introduce a different hypothesis, H_B, which introduces the specific relationship between parallax and magnitude shown in Fig. 1 (left panel). Although there is a relationship (limiting the probability of extreme values of D_2), the two quantities are independent. So again we have conditional independence, $P(D_1, D_2|H_B) = P(D_1|H_B)P(D_2|H_B)$, because given the hypothesis H_B, knowing D_2 tells us nothing additional about D_1.

Consider now a third hypothesis, H_A, shown in Fig. 1 (right panel), in which the parallax and magnitude are not independent. Clearly, knowledge of D_2 *does* now tell us something about (constrain) D_1, i.e. $P(D_1|D_2, H_A) \neq P(D_1|H_A)$ and hence $P(D_1|D_2, H_A) \neq P(D_1|H_A)P(D_1|H_B)$. So D_1 and D_2 are not conditionally independent given H_A.

We can of course generalize this and think of H not as a model, but some other measurement, parameter or piece of information. The point is that introducing information H may, but does not necessarily, make two measurements which are initially dependent conditionally independent (think of the rugby match example), or make two measurements which are initially independent conditionally dependent (think of starting with H_B then introducing H_A).

In a real situation we might choose to make D_1 and D_2 independent. But if the true measurements were actually dependent via a relationship like H_A, then we would make incorrect inferences. In such a case we should rather build a two-dimensional model of $P(D_1, D_2|H_A)$ and use this in the inference of the parameters of interest.

Some confusion about independence could be averted if we avoided speaking of "unconditional independence". *All* probabilities and knowledge are conditional on something. It helps considerably to identify what that something is.

4.3 Astrophysical situation

The Gaia Discrete Source Classifier (KS-019, CHL-004, CBJ-040, Bailer-Jones et al. 2008) will classify objects based on spectra, $D_{\rm spec}$, parallax/proper motion, $D_{\rm ast}$, Galactic coordinates, $D_{\rm pos}$, and apparent magnitude, D_G. We need to identify which of these, in practice, are "independent" and so can be modelled separately.

Unconditional on $D_{\rm spec}$ or $D_{\rm pos}$, but conditional on some sensible model for the universe H_U, $D_{\rm ast}$ and D_G are not independent, because H_U tells us that fainter objects tend to be more distant (analogous to the right panel of Fig. 1). So $P(D_{\rm ast}, D_G|H_U) \neq P(D_{\rm ast}|H_U)P(D_G|H_U)$. If we now knew that the object was a star, $C = C_{\rm star}$, then $D_{\rm ast}$ and D_G are still not independent conditioned on C, because fainter stars also tend to be more distant. So $P(D_{\rm ast}, D_G|H_U, C = C_{\rm star}) \neq P(D_{\rm ast}|H_U, C = C_{\rm star})P(D_G|H_U, C = C_{\rm star})$. On the other hand, if we knew that the object was a quasar, $C = C_{\rm quasar}$, then $D_{\rm ast}$ and D_G do become independent conditioned on C, because H_U tells us that all quasars have zero expected parallax and proper motion (any deviation just being noise). We see that the effect of conditioning on a variable depends not only on what the variable is but possibly also its value.

What about the independence of $D_{\rm spec}$ and $D_{\rm ast}$? Conditioned only on H_U then they are not strictly independent, because spikier spectra (which are more likely to be quasars) will generally have smaller values of the astrometry. But this is because H_U implicitly has information about what quasars look like. So it depends on the details of H_U. If we condition on $C = C_{\rm star}$, then it is fair to assume that they become independent. That is, given that the object is a star (plus all the general background information about astrophysical objects embodied in H_U), the astrometry tells us nothing about the spectrum and vice versa. Strictly one could claim that if the spectrum suggests a very cool dwarf, then this is more likely to be nearer, because such stars are faint and will only be visible by Gaia if nearby.

Generally, we should not assume independence of the data conditional only on background information such as H_U. But conditional on the class or parameters, we normally would assume conditional independence (the rubgy match example).

The complementary case of conditional independence of the stellar atmospheric *parameters*, given certain data, is discussed in section 2 of CBJ-056.

We examine some practical applications of combining data for Gaia classification in the technical note CBJ-037, and we use these methods in the classification of Blue Horizontal Branch stars based on photometry and magnitude in Smith et al. (2010). Bailer-Jones (2011b) combines colour, photometric, parallax and HRD prior information for estimating stellar parameters from Hipparcos/2MASS data using a Bayesian method (see CBJ-049 for an application of this method to Gaia).

References

Bailer-Jones C.A.L., 2011b, *Bayesian inference of stellar parameters and interstellar extinction using parallaxes and multiband photometry*, MNRAS 411, 425

Bailer-Jones C.A.L., 2011a, *Some more results on q-method. Probabilistic estimation of M_G, Z and log g*, GAIA-C8-TN-MPIA-CBJ-056

Bailer-Jones C.A.L., 2010, *Probabilistic combination of AP estimates based on spectra, astrometry and the HR Diagram with the aim of reducing degeneracy*, GAIA-C8-TN-MPIA-CBJ-049

Bailer-Jones C.A.L., 2008, *Developing further the Discrete Source Classifier*, GAIA-C8-TN-MPIA-CBJ-040

Bailer-Jones C.A.L., Smith K.S, 2008, *Multistage probabilistic classification*, GAIA-C8-TN-MPIA-CBJ-037

Bailer-Jones C.A.L., Smith K.S., Tiede C., Sordo R., Vallenari A., 2008, *Finding rare objects and building pure samples: probabilistic quasar classification from low-resolution Gaia spectra*, MNRAS 391, 1838

Liu C., et al., 2011, *CU8 Software design description for scientific algorithms*, GAIA-C8-SP-MPIA-CHL-004

Smith K.S, Bailer-Jones C.A.L., 2011, *DSC performance and status report*, GAIA-C8-TN-MPIA-KS-019

Smith K.W., Bailer-Jones C.A.L., Klement R., Xue X.X., 2010, *Photometric identification of Blue Horizontal Branch stars in SDSS*, A&A 522, A88

A Derivation of the normalization constant

For the sake of illustration we calculate the normalization constant a in equation 12 for a two-class problem (classes C and $\overline{C}$) for the general case where D_n does not necessarily include D_o. Normalization requires $P(C|D_A, D_n) + P(\overline{C}|D_A, D_n) = 1$, so

$$1 \;=\; a\left[P(C|D_A, D_o)\,\frac{P(C|D_n)}{P(C|D_o)} \;+\; P(\overline{C}|D_A, D_o)\,\frac{P(\overline{C}|D_n)}{P(\overline{C}|D_o)}\right] \tag{16}$$

For brevity we define $Q \equiv P(C|D_A, D_o)$, $R \equiv P(C|D_o)$ and $S \equiv P(C|D_n)$ and continue

$$\begin{aligned}\frac{1}{a} &= \frac{QS}{R} + \frac{(1-Q)(1-S)}{(1-R)} \\ &= \frac{QS - QRS + R - QR - RS + QRS}{R(1-R)} \\ &= \frac{R(1-Q-S) + QS}{R(1-R)}\quad .\end{aligned} \tag{17}$$

Substituting this into equation 12 gives

$$P(C|D_A, D_n) \;=\; a\frac{QS}{R} \;=\; \frac{QS(1-R)}{R(1-Q-S)+QS} \tag{18}$$