Conformal Prediction under Hypergraphical Models

Valentina Fedorova, Alex Gammerman,
Ilia Nouretdinov, and Vladimir Vovk

{valentina,ilia,alex,vovk}@cs.rhul.ac.uk

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #9

First posted July 3, 2013. Last revised January 18, 2022.

Project web site:
http://alrw.net

Abstract

Conformal predictors are usually defined and studied under the exchangeability assumption. However, their definition can be extended to a wide class of statistical models, called online compression models, while retaining their property of automatic validity. This paper is devoted to conformal prediction under hypergraphical models that are more specific than the exchangeability model. Namely, we define two natural classes of conformity measures for such hypergraphical models and study the corresponding conformal predictors empirically on benchmark LED data sets. Our experiments show that they are more efficient than conformal predictors that use only the exchangeability assumption.

Contents

1 Introduction

The method of conformal prediction was introduced and is usually used for producing valid prediction sets under the exchangeability assumption; the validity of the method means that the probability of making a mistake is equal to (or at least does not exceed) a prespecified significance level ([9], Chapter 2). However, the definition of conformal predictors can be easily extended to a wide class of statistical models, called online compression models (OCMs; [9], Chapter 8). OCMs compress data into a more or less compact summary, which is interpreted as the useful information in the data. With each "conformity measure", which, intuitively, estimates how well a new piece of data fits the summary, one can associate a conformal predictor, which still enjoys the property of automatic validity. Numerous machine learning algorithms have been used for designing efficient conformity measures (see, e.g., [9] and [2]), but this has been mostly done only for the exchangeability assumption.

This paper studies conformal prediction under the OCMs known as hypergraphical models ([9], Section 9.2). Such models describe relationships between data features. In the case where every feature is allowed to depend in any way on the rest of the features, the hypergraphical model becomes the exchangeability model. More specific hypergraphical models restrict the dependence in some way. Such restrictions are typical in modelling many real-world problems: for example, different symptoms might be assumed to be conditionally independent given the disease. A popular approach to such problems is to use Bayesian networks (see, e.g., [3]). The definition of Bayesian networks requires a specification of both the pattern of dependence between features and the distribution of the features. Usual methods guarantee a valid probabilistic outcome if the used distributions of features are correct. Several algorithms (see, e.g., [3], Chapter 9) are known for estimating the distribution of features; however, the accuracy of such approximations is a major concern in applying Bayesian networks. The conformal predictors constructed from hypergraphical OCMs use only the pattern of dependence between the features but do not involve their distribution. This makes conformal prediction based on hypergraphical models more robust and realistic than Bayesian networks. (The notion of a hypergraphical model can be regarded as more general than that of a Bayesian network: the standard algorithms in this area transform Bayesian networks into hypergraphical models by "marrying parents", forgetting the direction of the arrows, triangulation, and regarding the cliques of the resulting graph as the hyperedges; see, e.g., [3], Section 3.2.)

As far as we know, conformal prediction has been studied, apart from the exchangeability model and its variations, only for the Gauss linear model and Markov model (see [9], Chapter 8, and [5]). Hypergraphical OCMs have been used only in the context of Venn rather than conformal prediction (see [9], Chapter 9).

The rest of the paper is organised as follows. Section 2 formally defines hypergraphical OCMs and briefly reviews their basic properties. Section 3 describes the method of conformal prediction in the context of hypergraphi-

cal models and introduces two conformity measures for hypergraphical OCMs. Section 4 reports the performance of the corresponding conformal predictors on benchmark LED data sets. Section 5 describes the label conditional version of conformal predictors for hypergraphical OCMs, and also empirically studies class-wise validity and efficiency for these predictors. Section 6 concludes.

A weakness of using hypergraphical models different from the exchangeability model is that, even though they are much weaker than the corresponding Bayesian models, they still make much stronger assumptions about the data than the exchangeability model; for this reason we only experiment with an artificial data set in this paper (Section 4), as it is difficult to find real-world data sets for which even our assumptions would be completely realistic. However, one of our empirical findings in Section 4 is that the efficiency of our predictors does not suffer much when they are made valid under the exchangeability model while still using narrower hypergraphical models in their design. (Unfortunately, the situation changes if we insist on class-wise validity, as we will see in Section 5.)

2 Background

Consider two measurable spaces $\mathbf{X}$ and $\mathbf{Y}$; elements of $\mathbf{X}$ are called *objects* and elements of $\mathbf{Y}$ are called *labels*. Elements of the Cartesian product $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ are called *observations*. A *training sequence* is a sequence of observations $(z_1, \dots, z_l)$, where each observation $z_i = (x_i, y_i)$ consists of an object x_i and its label y_i. The general prediction problem considered in this paper is to predict the label for a new object given a training sequence. We focus on the case where $\mathbf{X}$ and $\mathbf{Y}$ are finite.

2.1 Hypergraphical structures

In this paper we assume that objects are structured, consisting of variables (representing features). Hypergraphical structures describe relationships between the variables. Formally a *hypergraphical structure*[1] consists of three elements $(V, \mathcal{E}, \Xi)$:

1. V is a finite set; its elements are called *variables*.
2. $\mathcal{E}$ is a finite collection of subsets of V whose union covers all variables: $\bigcup_{E \in \mathcal{E}} E = V$. Elements of $\mathcal{E}$ are called *clusters*.
3. Ξ is a function that maps each variable $v \in V$ into a finite set (of the values that v can take).

A *configuration* on a set $E \subseteq V$ (we are usually interested in the case where E is a cluster) is an assignment of values to the variables from E; let $\Xi(E)$ be the

[1] The name reflects the fact that the components $(V, \mathcal{E})$ form a hypergraph, where a hyperedge $E \in \mathcal{E}$ can connect more than two vertices.

set of all configurations on E. A *table*[2] on a set E is an assignment of natural numbers or zero to the configurations on E. The *size* of the table is the sum of values that it assigns to different configurations. A *table set* is a collection of tables on the clusters $\mathcal{E}$, one for each cluster $E \in \mathcal{E}$. The number assigned by a table set σ to a configuration on E is called its σ-*count*.

Intuitive examples of hypergraphical structures are given in Subsection 4.2 below.

2.2 Hypergraphical online compression models

The observation space $\mathbf{Z}$ associated with the hypergraphical structure is the set of all configurations on V. One of the variables in V is singled out as the *label variable*, and the configurations on the label variable are denoted $\mathbf{Y}$. All other variables are *object variables*, and the configurations on the object variables are denoted $\mathbf{X}$. Since $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$, this is a special case of the prediction setting described in Subsection 2.1.

An observation $z \in \mathbf{Z}$ *agrees* with a configuration on a set $E \subseteq V$ (or the configuration agrees with the observation) if the restriction $z|_E$ of z to the variables in E coincides with the configuration. A table set σ *generated* by a sequence of observations $(z_1, \ldots, z_n)$ assigns to each configuration on each cluster the number of observations in the sequence that agree with the configuration; the size of each table in σ will be equal to the number of observations in the sequence, and this number is called the *size* of the table set. Different sequences of observations can generate the same table set σ, and we denote $\#\sigma$ the number of different sequences generating σ. Whereas $\#\sigma > 0$ implies that the size of σ exists (i.e., all tables in σ have the same size), it is clear that the opposite implication is false in general.

The *hypergraphical online compression model* (HOCM) associated with the hypergraphical structure $(V, \mathcal{E}, \Xi)$ consists of five elements $(\Sigma, \square, \mathbf{Z}, F, B)$, where:

1. The *empty table set* $\square$ is the table set assigning 0 to each configuration.
2. The set Σ is defined by the conditions that $\square \in \Sigma$ and $\Sigma \setminus \{\square\}$ is the set of all table sets σ with $\#\sigma > 0$. The elements $\sigma \in \Sigma$ are called *summaries*.
3. The *forward function* $F(\sigma, z)$, where σ ranges over Σ and z over $\mathbf{Z}$, updates σ by adding 1 to the σ-count of each configuration which agrees with z.
4. The *backward kernel* B maps each $\sigma \in \Sigma \setminus \{\square\}$ to a probability distribution $B(\sigma)$ on $\Sigma \times \mathbf{Z}$ assigning the weight $\#(\sigma \downarrow z)/\#\sigma$ to each pair $(\sigma \downarrow z, z)$, where z is an observation such that, for all configurations which agree with z, the corresponding σ-counts are positive, and $\sigma \downarrow z$ is the table set obtained by subtracting 1 from the σ-counts of the configurations that agree with z. Notice that $B(\sigma)$ is indeed a probability distribution, and it is concentrated on the pairs $(\sigma \downarrow z, z)$ such that $F(\sigma \downarrow z, z) = \sigma$.

[2] Generally, a table assigns real numbers to configurations. In this paper we only consider *natural tables*, which assign natural numbers or zero to configurations, and omit "natural" for brevity.

We will use "hypergraphical models" as a general term for hypergraphical structures and HOCMs when no precision is required. When discussing hypergraphical models we will always assume that the observations $z_1, z_2, \ldots$ are produced independently from a probability distribution Q on $\mathbf{Z}$ that has a decomposition

$$Q(\{z\}) = \prod_{E \in \mathcal{E}} f_E(z|_E) \tag{1}$$

for some functions $f_E : \Xi(E) \to [0,1]$, $E \in \mathcal{E}$, where z is an observation and $z|_E$ its restriction to the variables in E.

2.3 Junction tree structures

An important type of hypergraphical structures is where clusters can be arranged into a "junction tree". For the corresponding HOCMs we will be able to describe efficient calculations of the backward kernels. If one wants to use the calculations for a structure that cannot be arranged into a junction tree it can be replaced by a more general junction tree structure before defining the HOCM.

Let (U, S) denote an undirected tree with U the set of vertices and S the set of edges. Then (U, S) is a *junction tree* for a hypergraphical structure $(V, \mathcal{E}, \Xi)$ if there exists a bijective mapping C from the set of vertices U of the tree to the set $\mathcal{E}$ of clusters of the hypergraphical structure that has the following property: $C_u \cap C_w \subseteq C_v$ whenever a vertex v lies on the path from a vertex u to a vertex w in the tree (we let C_x stand for $C(x)$). Not every hypergraphical structure admits a junction tree, of course: an example is a hypergraphical structure with three clusters whose intersection is empty but whose pairwise intersections are not. See, e.g., [3], Section 4.3, for further information on junction trees; intuitive examples of junction trees will be given in Section 4.

If $s = \{u, v\} \in S$ is an edge of the junction tree connecting vertices u and v then C_s stands for $C_u \cap C_v$. It is convenient to identify vertices u and edges s of the junction tree with the corresponding clusters C_u and sets C_s, respectively.

If $E_1 \subseteq E_2 \subseteq V$ and f is a table on E_2, the *marginalisation* of f to E_1 is the table f^* on E_1 assigning to each $a \in \Xi(E_1)$ the number $f^*(a) = \sum_b f(b)$, where b ranges over the configurations on E_2 such that $b|_{E_1} = a$. If σ is a summary then for $u \in U$ denote σ_u the table that σ assigns to C_u, and for $s = \{u, v\} \in S$ denote σ_s the marginalisation of σ_u (or σ_v) to C_s. We will use the shorthand $\sigma_u(z)$ for the number assigned to the restriction $z|_{C_u}$ by the table for the vertex u and $\sigma_s(z)$ for the number assigned to $z|_{C_s}$ by the marginal table for the edge s:

$$\sigma_u(z) := \sigma\left(z|_{C_u}\right), \quad \sigma_s(z) := \sigma\left(z|_{C_s}\right).$$

Consider the HOCM corresponding to the junction tree (U, S). We use the notation $P_\sigma(z)$ for the weight assigned by $B(\sigma)$ to $(\sigma \downarrow z, z)$. It has been proved ([9], Lemma 9.5) that

$$P_\sigma(z) = \frac{\prod_{u \in U} \sigma_u(z)}{n \prod_{s \in S} \sigma_s(z)}, \tag{2}$$

where n is the size of σ. If any of the factors in (2) is zero then the whole ratio is set to zero.

3 Conformal prediction for HOCM

Consider a training sequence $(z_1, \dots, z_l)$ and an HOCM $(\Sigma, \square, \mathbf{Z}, F, B)$. The goal is to predict the label for a new object x.

A *conformity measure* for the HOCM is a measurable function $A : \Sigma \times \mathbf{Z} \to \mathbb{R}$. The function assigns a *conformity score* $A(\sigma, z)$ to an observation z w.r. to a summary σ. Intuitively, the score reflects how typical it is to observe z after observing data summarized by σ.

For each $y \in \mathbf{Y}$ denote $\sigma^* \in \Sigma$ the table set generated by the sequence $(z_1, \dots, z_l, (x, y))$ (the dependence of σ^* on y is important although not reflected in our notation). For $z \in \mathbf{Z}$ such that $\sigma^* \downarrow z$ is defined denote the conformity scores as

$$\alpha_z := A(\sigma^* \downarrow z, z) \tag{3}$$

(notice that $\alpha_{(x,y)}$ is always defined). The *p-value* for y, denoted $p^{(y)}$, is defined by

$$p^{(y)} := \sum_{z:\alpha_z < \alpha_{(x,y)}} P_{\sigma^*}(z) + \theta \sum_{z:\alpha_z = \alpha_{(x,y)}} P_{\sigma^*}(z) \tag{4}$$

(cf. (8.4) in [9]), where $\theta \sim \mathbf{U}[0,1]$ is a random number drawn from the uniform distribution on $[0,1]$, $P_{\sigma^*}(z)$ is the backward kernel, as defined above, and the sums involve only those $z \in \mathbf{Z}$ for which α_z is defined. Then for a significance level ϵ the *hypergraphical conformal predictor* Γ based on A outputs the prediction set

$$\Gamma^\epsilon(z_1, \dots, z_l, x) := \{y \in \mathbf{Y} : p^{(y)} > \epsilon\}. \tag{5}$$

(Such randomized conformal predictors were referred to as "smoothed" in [9].)

We will describe two conformity measures for HOCMs in Subsection 3.1. These conformity measures optimise different criteria for the quality of conformal predictors. Subsection 3.3 will describe the criteria used in this paper.

The reader who is looking for an accessible and detailed description of conformal prediction can consult [7] (whose Section 2 gives a very simple example, albeit for a different, much simpler, online compression model).

3.1 Conformity measures for HOCM

Consider a summary σ and an observation (x, y). The *conditional probability conformity measure* is defined by

$$A(\sigma, (x, y)) := P_{\sigma^*}(y \mid x) := \frac{P_{\sigma^*}((x, y))}{\sum_{y' \in \mathbf{Y}} P_{\sigma^*}((x, y'))}, \tag{6}$$

where $\sigma^* := F(\sigma, (x, y))$ and $P_{\sigma^*}((x, y))$ is the backward kernel. In other words, $A(\sigma, (x, y))$ is the conditional probability $P_{\sigma^*}(y \mid x)$ of y given x under P_{σ^*}. The conditional probability $P_{\sigma^*}(y \mid x)$ can be easily computed using (2).

Define the *predictability* of an object $x \in \mathbf{X}$ as

$$f(x) := \max_{y \in \mathbf{Y}} P_{\sigma^*}(y \mid x), \tag{7}$$

the maximum of conditional probabilities. If the predictability of an object is close to 1 then the object is "easily predictable". Fix a *choice function* $\hat{y} : \mathbf{X} \to \mathbf{Y}$ such that

$$\forall x \in \mathbf{X} : f(x) = P_{\sigma^*}(\hat{y}(x) \mid x).$$

This function maps each object x to one of the labels at which the maximum in (7) is attained. The *signed predictability conformity measure* is defined by

$$A(\sigma, (x, y)) := \begin{cases} f(x) & \text{if } y = \hat{y}(x) \\ -f(x) & \text{otherwise.} \end{cases} \tag{8}$$

3.2 Computational efficiency of conformal prediction for HOCM

In this paper we study the performance of conformal predictors in the online prediction protocol (Protocol 1). The prediction sets output by Predictor are computed using a conformal predictor Γ: $\Gamma^\epsilon_n := \Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$ for some $\epsilon \in (0, 1)$ (or for a finite range of ϵ). In this section we will discuss the computational efficiency of Γ in this protocol.

Protocol 1 Online prediction protocol

for $n = 1, 2, \dots$
Reality outputs $x_n \in \mathbf{X}$
Predictor outputs $\Gamma^\epsilon_n \subseteq \mathbf{Y}$ for $\epsilon \in (0, 1)$
Reality outputs $y_n \in \mathbf{Y}$

Let us assume that the HOCM $(\Sigma, \square, \mathbf{Z}, F, B)$ and the finite range of ϵ are fixed (for concreteness, the reader can think of the first example of Subsection 4.2). We will see that in this case the computations at each step of the online prediction protocol can be carried out in constant time, $O(1)$, for both the conditional probability conformity measure (6) and the signed predictability conformity measure (8); it will be clear that this is true for a very wide range of conformity measures.

Indeed, let σ_n be the summary of the first n observations $z_1, \dots, z_n$. According to item (3) of the definition of HOCMs, updating σ_n (i.e., computing σ_n from σ_{n-1} and z_n when $n > 0$) can be done in constant time. Given σ_{n-1}, x_n, and a postulated label y (of which there are a fixed finite number), we can compute the probability measure P_{σ^*} defined by (2) for the summary $\sigma^* := F(\sigma_{n-1}, (x_n, y))$ in constant time. All conformity scores (6) and (8) can now be computed in constant time. Finally, the p-values (4) and prediction set (5) can be computed in constant time.

3.3 Criteria for the quality of conformal prediction

Remember that we are working in the online prediction protocol, given as Protocol 1. Reality generates observations (x_n, y_n) from a probability distribution Q satisfying (1) for some hypergraphical structure. Predictor uses a conformal predictor Γ to output the prediction set $\Gamma^\epsilon_n := \Gamma^\epsilon(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$ at each significance level ϵ. (We are no longer interested in the computational efficiency, so we imagine that Predictor outputs Γ^ϵ_n for all $\epsilon \in (0,1)$; in practice, we can use either a finite range of ϵ or values such as unconf_n below which do not depend on ϵ and are efficiently computable.)

Two important properties of conformal predictors are their validity and efficiency; the first is achieved automatically and the second is enjoyed by different conformal predictors to a different degree. Predictor *makes an error* at step n if y_n is not in Γ^ϵ_n. The validity of conformal predictors means that, for any significance level ϵ, the probability of error $y_n \notin \Gamma^\epsilon_n$ is equal to ϵ. It has been proved that conformal predictors are automatically valid under their models ([9], Theorem 8.1). In this paper we study problems where the hypergraphical model used for computing the p-values is known to be correct; therefore, the predictions will always be valid, and there is no need to test validity experimentally.

The efficiency of valid predictions can be measured in different ways. The standard way is to count the *number of multiple predictions* Mult^ϵ_n over the first n steps defined by

$$\text{mult}^\epsilon_n := \begin{cases} 1 & \text{if } |\Gamma^\epsilon_n| > 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \text{Mult}^\epsilon_n := \sum_{i=1}^{n} \text{mult}^\epsilon_i$$

at each significance level $\epsilon \in (0,1)$ (cf. [9], Chapter 3). Another way is to report the *cumulative observed excess of prediction sets*

$$\text{OE}^\epsilon_n := \sum_{i=1}^{n} |\Gamma^\epsilon_i \setminus \{y_i\}| \tag{9}$$

at each significance level $\epsilon \in (0,1)$. (The observed excess of a prediction is the number of false labels included in the prediction set.) We will also consider two ways to measure the efficiency of conformal predictors that do not depend on the significance level. Let $p^{(y)}_n$, $y \in \mathbf{Y}$, be the p-values (4) used by the conformal predictor for computing the prediction set Γ^ϵ_n at the nth step of the online prediction protocol. The *cumulative unconfidence* Unconf_n over the first n steps is defined by

$$\text{unconf}_n := \inf\{\epsilon : |\Gamma^\epsilon_n| \le 1\} \qquad \text{and} \qquad \text{Unconf}_n := \sum_{i=1}^{n} \text{unconf}_i;$$

the *unconfidence* unconf_n at step n can be equivalently defined as the second largest p-value among $p^{(y)}_n$, $y \in \mathbf{Y}$. (Unconfidence is a trivial modification of

the standard notion of confidence: see [9], (3.66).) Finally, the efficiency can be measured by the *cumulative observed fuzziness*

$$\mathrm{OF}_n := \sum_{i=1}^{n} \sum_{y \in \mathbf{Y}: y \neq y_i} p_i^{(y)}. \tag{10}$$

(The observed fuzziness at step n is the sum of the p-values apart from that for the true label.) All four criteria work in the same direction: the smaller the better. As already mentioned, the number of multiple predictions is a standard criterion; the three other criteria are first used in this paper, in our other recent paper [8], and in Johansson et al.'s [6] (we learned about the last paper only after the conference version [4] of this paper had been published).

In our experiments we will use the following more intuitive versions of the first two criteria: the *percentage of multiple predictions* $\mathrm{Mult}_n^\epsilon/n$ and the *average observed excess of predictions* OE_n^ϵ/n; we would like them to be close to 0 for small significance levels.

3.4 Conformity measures and criteria of efficiency

It can be shown that, in a wide range of situations:

- the signed predictability conformity measure is optimal in the sense of Mult_n^ϵ and in the sense of Unconf_n;
- the conditional probability conformity measure is optimal in the sense of OE_n^ϵ and in the sense of OF_n.

These statements are formalized in an asymptotic setting and proved in [8], but only in the case of the exchangeability model. However, in this subsection we will see that this observation extends to any hypergraphical models admitting a junction tree.

Intuitively, the asymptotic setting of [8] corresponds to the limiting case of infinitely long training and test sequences. This is formalized by assuming that the prediction algorithm is directly given the data-generating probability distribution Q on $\mathbf{Z}$ instead of being given training and test sequences. Conformity measures are replaced by *idealized conformity measures*: functions $A(Q, z)$ of $Q \in \mathcal{P}(\mathbf{Z})$ and $z \in \mathbf{Z}$ (where $\mathcal{P}(\mathbf{Z})$ is the set of all probability measures on $\mathbf{Z}$). The *idealized conformal predictor* corresponding to A outputs the following prediction set $\Gamma^\epsilon(x)$ for each object $x \in \mathbf{X}$ and each significance level $\epsilon \in (0, 1)$. For each potential label $y \in \mathbf{Y}$ for x define the corresponding p-value as

$$p^y = p(x, y) := Q\{z \in \mathbf{Z} \mid A(z) < A((x, y))\} + \theta Q\{z \in \mathbf{Z} \mid A(z) = A((x, y))\}, \tag{11}$$

where $\theta \sim \mathbf{U}[0, 1]$; the prediction set is

$$\Gamma^\epsilon(x) := \{y \in \mathbf{Y} \mid p(x, y) > \epsilon\}. \tag{12}$$

Equations (11) and (12) are the idealized versions of (4) and (5), respectively. We can also define the idealized version

$$A(Q,(x,y)) := Q(y \mid x) := \frac{Q\left(\{(x,y)\}\right)}{Q\left(\{x\} \times \mathbf{Y}\right)}$$

of the conditional probability conformity measure (6) and the idealized version

$$A(Q,(x,y)) := \begin{cases} f(x) & \text{if } y = \hat{y}(x) \\ -f(x) & \text{otherwise.} \end{cases}$$

of the signed predictability conformity measure (8), where (7) is redefined as

$$f(x) := \max_{y \in \mathbf{Y}} Q(y \mid x).$$

Finally, we can define idealized version of the criteria of efficiency; we will do so only for the criteria (9) and (10). Let us write $\Gamma^\epsilon_A(x)$ for the $\Gamma^\epsilon(x)$ in (12) and $p_A(x,y)$ for the $p(x,y)$ in (11) to indicate the dependence on the choice of the idealized conformity measure A. An idealized conformity measure A is:

- *OF-optimal* if, for any idealized conformity measure B,

$$\mathbb{E}_{(x,y),\theta} \sum_{y' \neq y} p_A(x,y') \leq \mathbb{E}_{(x,y),\theta} \sum_{y' \neq y} p_B(x,y'), \tag{13}$$

  where the notation $\mathbb{E}_{(x,y),\theta}$ refers to the expected value when $(x,y) \sim Q$ and $\theta \sim \mathbf{U}[0,1]$ independently;

- *OE-optimal* if, for any idealized conformity measure B and any significance level ϵ,

$$\mathbb{E}_{(x,y),\theta} \left|\Gamma^\epsilon_A(x) \setminus \{y\}\right| \leq \mathbb{E}_{(x,y),\theta} \left|\Gamma^\epsilon_B(x) \setminus \{y\}\right|. \tag{14}$$

Now we can state the result about the OF- and OE-optimality of the conditional probability conformity measure (Theorem 1 in [8]). We say that an idealized conformity measure A is a *refinement* of an idealized conformity measure B if

$$B(z_1) < B(z_2) \Longrightarrow A(z_1) < A(z_2)$$

for all $z_1, z_2 \in \mathbf{Z}$. Let $\mathcal{R}(\mathrm{CP})$ be the set of all refinements of the conditional probability idealized conformity measure. If C is a criterion of efficiency (OF or OE), we let $\mathcal{O}(C)$ stand for the set of all C-optimal idealized conformity measures. Theorem 1 in [8] says that

$$\mathcal{O}(\mathrm{OF}) = \mathcal{O}(\mathrm{\text{Œ}}) = \mathcal{R}(\mathrm{CP}). \tag{15}$$

In [8], (15) was considered to be an asymptotic formalization of the optimality of the conditional probability conformity measure under the exchangeability assumption. Let us check that (15) still formalizes the optimality of the conditional probability conformity measure under any hypergraphical model admitting a junction tree. As usual, we suppose that the data-generating distribution

Q has a decomposition (1). If a sequence of observations $z_1, z_2, \ldots$ is generated from Q independently and σ^* is the summary of $(z_1, \ldots, z_l, (x, y))$, we can see from (2) that $P_{\sigma^*} \to Q$ almost surely uniformly in (x, y) as $l \to \infty$ (remember that $\mathbf{Z}$ is finite, so that the notion of convergence for measures is unambiguous and the uniformity is automatic). To check this convergence, direct the junction tree designating an arbitrary vertex as the root $\square$ and directing all edges away from the root; we can then rewrite (2) as

$$P_\sigma(z) = \frac{\sigma_\square(z)}{n} \prod_{u \in U \setminus \{\square\}} \frac{\sigma_u(z)}{\sigma_{u'}(z)} \tag{16}$$

(cf. [9], (9.6)), where u' is the edge between u and its parent; and we can then see that each fraction in (16) converges to the corresponding conditional probability. Therefore, all conditional probability and signed predictability conformity scores converge to their idealized versions almost surely. On the other hand, using cumulative observed excess and fuzziness are equivalent to using average observed excess and fuzziness, which converge to expected observed excess and fuzziness by the strong law of large numbers. Therefore, in the limit the efficiency criteria (9) and (10) approach the idealized criteria (14) and (13). Formally, we can say the following. If a sequence of observations $z_n = (x_n, y_n)$, $n = 1, 2, \ldots$, is generated from a decomposable Q independently,

$$\frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_\theta \sum_{y \neq y_n} p^y_{\mathrm{CP},n} \le \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_\theta \sum_{y \neq y_n} p^y_{B,n} + o(1) \quad \text{a.s. as } N \to \infty, \tag{17}$$

where $p^y_{A,n}$ is the p-value defined by (4) for the training sequence $(z_1, \ldots, z_{n-1})$, test object x_n, and conformity measure A, CP is the conditional probability conformity measure, and B is any other conformity measure. Equation (17) is a more realistic (albeit still asymptotic) counterpart of the OF criterion (13). And a more realistic (also asymptotic) counterpart of the OE criterion (14) is

$$\frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_\theta \left| \Gamma^{\epsilon'}_{\mathrm{CP},n} \setminus \{y_n\} \right| \le \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_\theta \left| \Gamma^{\epsilon}_{B,n} \setminus \{y_n\} \right| + o(1) \quad \text{a.s. as } N \to \infty$$

for any confidence level ϵ and any $\epsilon' \in (\epsilon, 1)$ (arbitrarily close to ϵ), where $\Gamma^\epsilon_{A,n}$ is defined by (5) for the training sequence $(z_1, \ldots, z_{n-1})$, test object x_n, and conformity measure A.

We discussed in detail only the conditional probability conformity measure; the signed predictability conformity measure (which is also considered in [8]) is treated similarly (but more messily). The criteria of efficiency under which the conditional probability conformity measure becomes optimal are called probabilistic in [8], and that paper argues that probabilistic criteria have important advantages over the more traditional efficiency criteria based on the number of multiple predictions or cumulative (or average) (un)confidence.

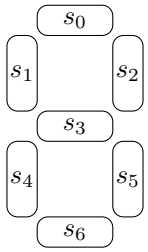

Figure 1: The seven-segment display.

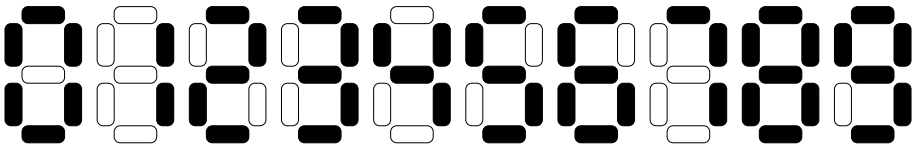

Figure 2: Ideal LED images.

4 Experimental Results

This section reports experimental results comparing various versions of conformal predictors using hypergraphical models. We use an artificial data set since it is not easy to find a real-world data set in which the assumptions of a non-trivial hypergraphical model were clearly satisfied.

4.1 LED data sets

For our experiments we use benchmark LED data sets generated by a program from the UCI repository [1]. The problem is to predict a digit from an image in the seven-segment display shown in Figure 1.

Figure 2 shows examples of objects in the data set (these are the ten "ideal images" of digits; there are also digits corrupted by noise). The seven LEDs (light emitting diodes) can be lit in different combinations to represent a digit from 0 to 9. The program generates observations with noise. There is an ideal image for each digit, as shown in Figure 2. An observation has seven binary attributes $s_0, \ldots, s_6$ (s_i is 1 if the ith LED is lit) and a label c, which is a decimal digit. The program randomly chooses a label (0 to 9 with equal probabilities), inverts each of the attributes of its ideal image with probability $p_{\text{noise}} := 1\%$ independently, and adds the noisy image and its label to the data set.

Let $(S_0, \ldots, S_6, C)$ be the vector of random variables corresponding to the attributes and the label, and let $(s_0, \ldots, s_6, c)$ be an observation. According to the data-generating mechanism the probability of the observation decomposes

as

$$Q\left(\{(s_0,\ldots,s_6,c)\}\right) = Q_7\left(C=c\right)\cdot\prod_{i=0}^{6} Q_i\left(S_i=s_i \mid C=c\right), \tag{18}$$

where Q_7 is the uniform distribution on the decimal digits and

$$Q_i\left(S_i=s_i \mid C=c\right) := \begin{cases} 1-p_{\text{noise}} & \text{if } s_i = s_i^c \\ p_{\text{noise}} & \text{otherwise,} \end{cases} \qquad i=0,\ldots,6, \tag{19}$$

$(s_0^c,\ldots,s_6^c)$ being the attributes of the ideal image for the label c. As usual, observations are generated independently.

4.2 Hypergraphical assumptions for LED data sets

We consider two hypergraphical models that agree with the decomposition (18). These models make different assumptions about the pattern of dependence between the attributes and the label; they do not depend on a particular probability of noise p_{noise} or the fact that the same value of p_{noise} is used for all LEDs. For both hypergraphical structures the set of variables is $V := \{s_0,\ldots,s_6,c\}$.

Nontrivial hypergraphical model

Consider the hypergraphical structure with the clusters

$$\mathcal{E} := \{\{s_i,c\} : i=0,\ldots,6\}.$$

A junction tree for this hypergraphical structure can be defined as a chain with vertices $U := \{u_i : i=0,\ldots,6\}$ and the bijection $C_{u_i} := \{s_i,c\}$. By saying that U is a chain we mean that there are edges connecting vertices 0 and 1, 1 and 2, 2 and 3, 3 and 4, 4 and 5, and 5 and 6 (and these are the only edges in the tree). It is clear that this is a junction tree and that $C_s = \{c\}$ for each edge s. It is also clear from (18) that the assumption (1) is satisfied; e.g., we can set

$$\begin{aligned} f_{\{s_0,c\}}\left(s_0,c\right) &:= Q_7\left(C=c\right)\cdot Q_0\left(S_0=s_0 \mid C=c\right); \\ f_{\{s_i,c\}}\left(s_i,c\right) &:= Q_i\left(S_i=s_i \mid C=c\right), \quad i=0,\ldots,6. \end{aligned}$$

Exchangeability model

The hypergraphical model with no information about the pattern of dependence between the attributes and the label is the exchangeability model. The corresponding hypergraphical structure has one cluster, $\mathcal{E} := \{V\}$. The junction tree is the one vertex associated with V and no edges.

4.3 Experiments

For our experiments we create a LED data set with $10,000$ observations. The data are generated according to the model (18) with the probability of noise

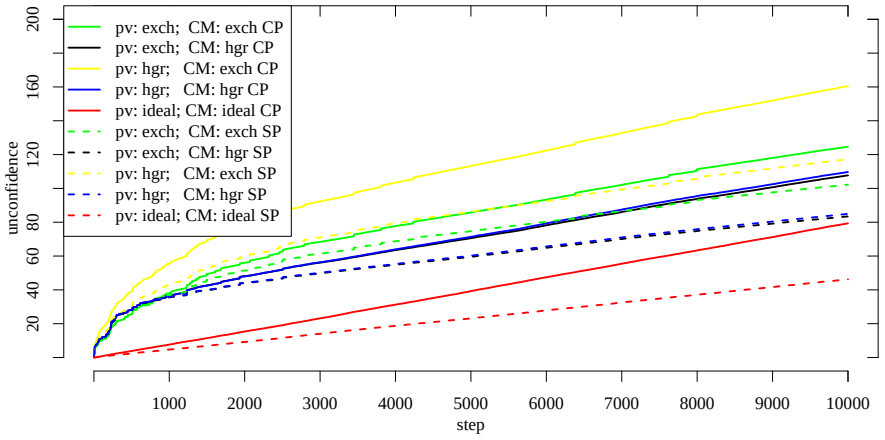

Figure 3: Cumulative unconfidence for online predictions. The results are for the LED data set with 1% of noise and 10,000 observations.

Table 1: The final values of the cumulative unconfidence in Figure 3 for the black and blue graphs.

Seed (10^4)	0	1	...	99	Average	St. dev.
pv: exch; CM: hgr CP	107.69	108.03	...	106.96	106.23	9.85
pv: hgr; CM: hgr CP	109.68	107.80	...	107.80	105.83	9.82
pv: exch; CM: hgr SP	83.40	90.26	...	89.09	82.19	7.07
pv: hgr; CM: hgr SP	84.89	90.56	...	89.45	82.39	6.81

$p_{\text{noise}} = 1\%$. Both for data generation and data processing, we set the seed of the pseudorandom number generator to 0. The text below assumes that the reader can see Figures 3–6 in colour; the colours become different shades of grey in black-and-white. We hope our descriptions will be detailed enough for the reader to identify the most important graphs unambiguously.

Each of the figures corresponds to an efficiency criterion for conformal predictors; namely, Figure 3 plots Unconf_n versus $n = 1, \dots, 10000$ in the online prediction protocol, Figure 4 plots OF_n versus $n = 1, \dots, 10000$, Figure 5 plots $\text{Mult}^\epsilon_{10000}/10000$ (the percentage of multiple predictions) versus $\epsilon \in [0, 0.05]$, and Figure 6 plots $\text{OE}^\epsilon_{10000}/10000$ (the average excess of predictions) versus $\epsilon \in [0, 0.05]$. We consider two conformity measures (CM): the conditional probability (CP) conformity measure (6) and the signed predictability (SP) conformity measure (8). The graphs corresponding to the former are represented in our plots as solid lines, and the graphs corresponding to the latter are represented as dashed lines.

Two of the plots in each figure correspond to idealized predictors and are

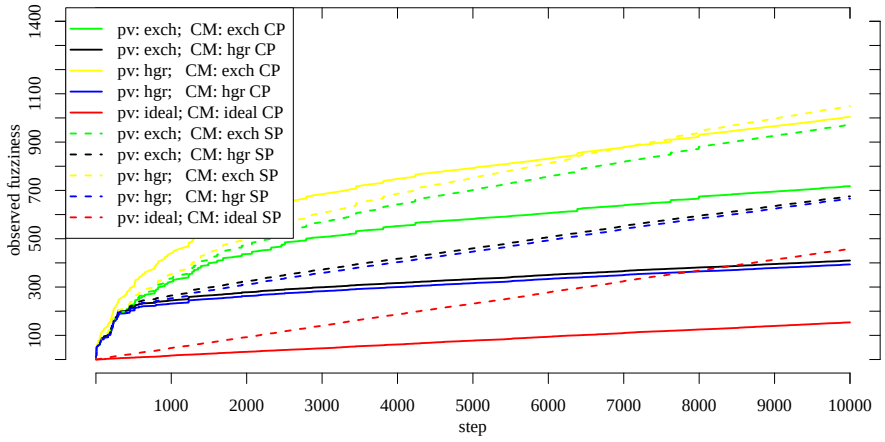

Figure 4: Cumulative observed fuzziness for online predictions. The results are for the LED data set with 1% of noise and 10,000 observations.

Table 2: The final values of the observed fuzziness in Figure 4 for the black and blue graphs.

Seed (10^4)	0	1	...	99	Average	St. dev.
pv: exch; CM: hgr CP	409.6	422.2	...	402.7	407.0	24.75
pv: hgr; CM: hgr CP	393.6	402.2	...	385.2	384.4	23.99
pv: exch; CM: hgr SP	675.4	747.0	...	717.1	676.1	55.87
pv: hgr; CM: hgr SP	666.1	729.8	...	701.2	657.2	53.90

drawn only for comparison, representing an unachievable ideal goal. In the idealized case we know the true distribution for the data (given by (18), (19), and $p_{\text{noise}} = 1\%$). The true distribution is used instead of the backward kernel P_{σ^*} in both (4) and (6) for the CP conformity measure and in both (4) and (8) for the SP conformity measure. It gives us the ideal results (the two red lines in our plots) for the two conformity measures, CP and SP. At least one of them gives the best results in each of the figures (remember that for all our criteria the lower the better).

For each of the two conformity measures we also consider four realistic predictors (which are conformal predictors, unlike the idealized ones). The *pure hypergraphical conformal predictor* (represented by blue lines in our plots) is obtained using the nontrivial hypergraphical model both when computing p-values (see (4)) and when computing the conformity measure ((6) in the case of CP and (8) in the case of SP). Analogously we use the exchangeability model to obtain the *pure exchangeability conformal predictor* (green lines in our plots). The two *mixed conformal predictors* (black and yellow lines) are obtained when

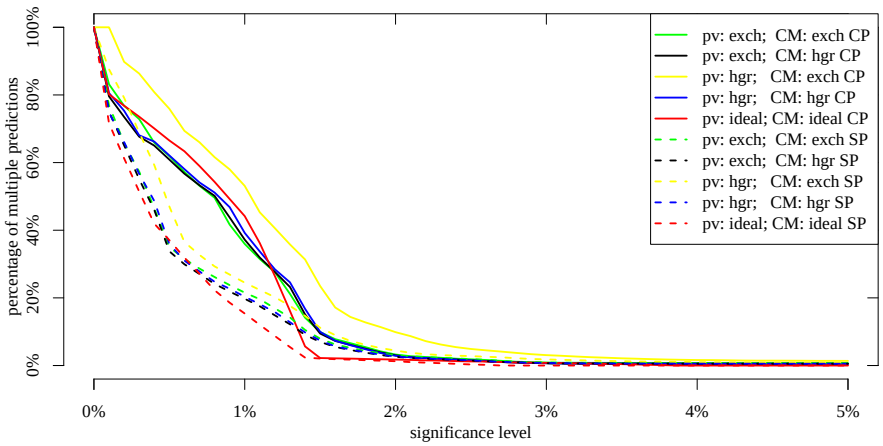

Figure 5: The final percentage of multiple predictions for significance levels between 0% and 5%. The results are for the LED data set with 1% of noise and 10,000 observations.

Table 3: The final percentage of multiple predictions in Figure 5 for the significance level 1% and for the black and blue graphs.

Seed (10^4)	0	1	...	99	Average	St. dev.
pv: exch; CM: hgr CP	0.3720	0.4046	...	0.4109	0.3812	0.0905
pv: hgr; CM: hgr CP	0.3920	0.4047	...	0.4128	0.3815	0.0896
pv: exch; CM: hgr SP	0.1972	0.2425	...	0.2478	0.1919	0.0516
pv: hgr; CM: hgr SP	0.2034	0.2437	...	0.2502	0.1962	0.0489

we use different models to compute the p-values and the conformity scores.

The intuition behind the pure and mixed conformal predictors can be explained using the distinction between hard and soft models made earlier in [10]. The model used when computing the p-values (see (4)) is the hard model; the validity of the conformal predictor depends on it. The model used when computing conformity scores (see (6) and (8)) is the soft model; when it is violated, validity is not affected, although efficiency can suffer. The true probability distribution (18) conforms to both the exchangeability model and the nontrivial hypergraphical model; therefore, all four conformal predictors are automatically valid, and we study only their efficiency. (In the context of this paper, it is obvious that the exchangeability model is more general than the nontrivial hypergraphical model, but we can also apply the criterion given in [9], Proposition 9.2.)

In the legends of Figures 3–6, the hard model used is indicated after "pv" (the way of computing the p-values), and the soft model used is indicated after "CM" (the conformity measure); "exch" refers to the exchangeability model,

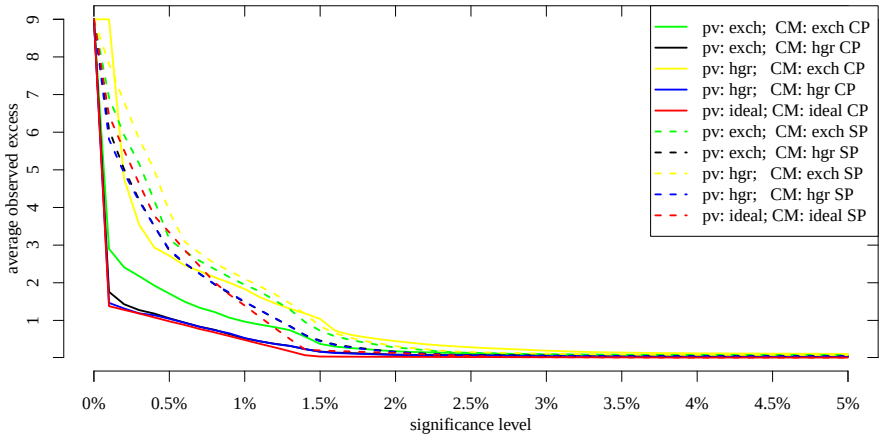

Figure 6: The final average observed excess for significance levels between 0% and 5%. The results are for the LED data set with 1% of noise and 10,000 observations.

Table 4: The final average observed excess in Figure 6 for the significance level 1% and for the black and blue graphs.

Seed (10^4)	0	1	...	99	Average	St. dev.
pv: exch; CM: hgr CP	0.5218	0.5115	...	0.5197	0.5451	0.1237
pv: hgr; CM: hgr CP	0.5299	0.4868	...	0.5025	0.5228	0.1216
pv: exch; CM: hgr SP	1.4870	1.7030	...	1.6648	1.4147	0.3598
pv: hgr; CM: hgr SP	1.4959	1.6327	...	1.6359	1.3806	0.3432

and "hgr" refers to the nontrivial hypergraphical model.

The most interesting graphs in Figures 3–6 are the black ones, corresponding to the exchangeability model as the hard model and the nontrivial hypergraphical model as the soft model. The performance of the corresponding conformal predictors is typically better than, or at least close to, the performance of any of the remaining realistic predictors. The fact that the validity of these conformal predictors only depends on the exchangeability assumption makes them particularly valuable. The yellow graphs correspond to the nontrivial hypergraphical model as the hard model and the exchangeability model as the soft model; the performance of the corresponding conformal predictors (rather inane, since it does not make sense for the hard model to be more restrictive than the soft model) is very poor in our experiments.

Now we will comment on each of the figures, and the corresponding tables, separately. In the case of the figures, the only available results are for the seed 0 of the pseudorandom number generator, but the corresponding tables and our

experiments not included in the paper confirm that our conclusions apply to other seeds as well.

Figure 3 shows the cumulative unconfidence Unconf_n, and so the right conformity measure to use is SP, as discussed at the end of Section 3; and indeed, all SP graphs lie below their CP counterparts. The two bottom graphs are the ones corresponding to idealized predictors; the graph corresponding to the CP idealized predictor, however, has a suboptimal slope. Of the realistic predictors, the lowest graph is the black SP one (but the blue SP graph, corresponding to the pure hypergraphical conformal predictor, is very close).

Table 1 shows the final values of the cumulative unconfidence in Figure 3 for the four most important graphs (two black and two blue) for several seeds. The values of the seed are given in the units of $10,000$ (so that 0 stands for 0, 1 for $10,000$, 2 for $20,000$, etc.), which is the minimal step to ensure that different experiments are based on completely different pseudorandom numbers (when the seed is initialized to n, the successive calls to the R pseudorandom number generator produce the pseudorandom numbers corresponding to the seeds n, $n+1$, $n+2$, etc.); the "10^4" in parentheses serves as a reminder of this. The last two columns of this and other tables give aggregate values: column "Average" gives the average of all the 100 values for the seeds 0–99, and column "St. dev." gives the standard estimate of the standard deviation computed from those 100 values (namely, the square root of the standard unbiased estimate of the variance). The table confirms that each black graph is very close to the corresponding blue graph on average (see the penultimate column), but the accuracy of our experiments is insufficient to say which tends to be lower: see the last column (to obtain an estimate of the standard deviation of the average, the value given in the last column should be divided by 10).

Figure 4 shows the cumulative observed fuzziness OF_n. For this criterion the predictors based on the CP conformity measure outperform the predictors based on the SP conformity measure (the solid lines are below the dashed lines of the same colour), as expected. The bottom graph corresponds to the idealized CP predictor; the idealized SP predictor is the second best most of the time, but at the end it is overtaken by the black and blue graphs corresponding to the conformal predictors based on the CP conformity measure using the nontrivial hypergraphical model. The black and blue graphs are very close; the blue one is slightly lower but the conformal predictor corresponding to the black one still appears preferable as its validity only depends on the weaker exchangeability assumption. Table 2 confirms that the black and blue graphs are close to each other on average, although there is a clear tendency for the blue ones to be lower.

Figure 5 shows the percentage of multiple predictions after observing $10,000$ observations as function of the significance level. For small significance levels the percentage of the multiple predictions is smaller for the predictors based on the SP conformity measure, again as expected. The performance of the conformal predictor corresponding to the black SP graph is again remarkably good, better than that of any other realistic predictor, although very close to the blue SP graph. According to Table 3, the accuracy of our experiments is insufficient to

tell whether the two blue graphs tend to be lower than the corresponding black ones at the significance level 1% for our data-generating mechanism.

Figure 6 shows the average observed excess of predictions after observing $10,000$ observations as function of the significance level. For small significance levels the predictors based on the CP conformity measure perform better, again confirming the theoretical results mentioned earlier. The black CP graph is very close to the blue CP graph, corresponding to the pure hypergraphical predictor, except for very low significance levels when the average excess exceeds 1. The closeness at the significance level 1% is confirmed by Table 4.

5 Label Conditional Conformal Prediction for HOCM

The usual notion of validity for conformal predictors is unconditional; the overall probability of error being equal to the significance level does not prevent the probability of error for different classes (such as 0s, 1s, etc. in the case of LED data sets) being different from ϵ, as long as the average probability over all classes remains ϵ. This section studies, theoretically and experimentally, label conditional conformal prediction under hypergraphical models, which achieves class-wise validity. We start with the formal definition of label conditional conformal predictors based on hypergraphical models, and follow by an empirical study of these predictors using an LED data set.

5.1 Definition and properties of label conditional conformal prediction

In general, the observations can be divided in a natural way into a finite number of categories (for example, each category can corresponds to a label, or to a kind of objects). We say that a conformal predictor is *category-wise valid* if for any significance level $\epsilon \in (0,1)$ the conditional probability of error given the test observation's category is ϵ. The automatic validity of conformal predictors (discussed in Subsection 3.3 above) does not guarantee their category-wise validity: for some categories error probabilities can be higher than the significance level, which could be balanced by lower error probabilities for other categories. A modification of conformal predictors that achieve category-wise validity, called Mondrian conformal predictors, were introduced in [9], Section 4.5, under the exchangeability assumption. This section studies Mondrian conformal predictors under hypergraphical models focusing on the categories corresponding to labels; the corresponding categories are called classes, as usual, and the corresponding Mondrian conformal predictors are called label conditional conformal predictors.

Formally, *hypergraphical label conditional conformal predictors* are defined in the same way as hypergraphical conformal predictors in Section 3 (see (3)–(5))

except that the definition (4) of p-values is modified as follows:

$$p^y := \frac{\sum_{(x',y')\in\mathbf{Z}:y'=y,\alpha_{(x',y')}<\alpha_{(x,y)}} P_{\sigma^*}((x',y'))}{\sum_{(x',y')\in\mathbf{Z}:y'=y} P_{\sigma^*}((x',y'))} + \theta\frac{\sum_{(x',y')\in\mathbf{Z}:y'=y,\alpha_{(x',y')}=\alpha_{(x,y)}} P_{\sigma^*}((x',y'))}{\sum_{(x',y')\in\mathbf{Z}:y'=y} P_{\sigma^*}((x',y'))}, \quad (20)$$

where, as usual, the sums involve only those $(x', y') \in \mathbf{Z}$ for which $\alpha_{(x',y')}$ is defined. Using these p-values instead of (4), hypergraphical label conditional conformal predictors are defined analogously to the hypergraphical conformal predictions in Section 3; we will sometimes refer to the latter as "unconditional" conformal predictors. To summarize, the conformity scores are defined by (3), the p-values are defined by (20), and the prediction sets by (5).

As in the unconditional case, we can use both the conditional probability conformity measure (6) and the signed predictability conformity measure (8) when computing the conformity scores (3). In this section we will only consider the former. In the label conditional case it is still true that the conditional probability conformity measure is optimal in the sense of OE^ϵ_n and in the sense of OF_n: see [11].

5.2 Empirical study

This subsection studies the performance of unconditional and label conditional conformal predictors under hypergraphical models. First we look at the class-wise validity of these predictors and then we compare their efficiency.

As before, the LED data set that we use consists of $10,000$ observations generated with $p_{\text{noise}} = 1\%$. The results presented in this section are for the seed 0 of the pseudorandom number generator, for both the data generation and data processing programs. The two hypergraphical models were described in Subsection 4.2. Predictors for these experiments are based on the hypergraphical CP conformity measure (6).

Class-wise validity

In our first experiment we assess the final percentage of errors within different classes. For each of the ten classes corresponding to labels $y \in \{0, \dots, 9\}$ and at each significance level $\epsilon \in (0, 1)$ the final percentage of errors is calculated by

$$\mathrm{Err}^{y,\epsilon} := \frac{|\{i = 1, \dots, 10000 \mid y_i = y, y_i \notin \Gamma^\epsilon_i\}|}{|\{i = 1, \dots, 10000 \mid y_i = y\}|}. \quad (21)$$

We study four conformal predictors: the pure exchangeability conformal predictor (the exchangeability model is used for the p-values (4) and for the conformity scores (6)), the pure hypergraphical conformal predictor (the nontrivial hypergraphical model is used in (4) and (6)), the *pure exchangeability label conditional conformal predictor* (the exchangeability model is used in (20) and in (6)), and

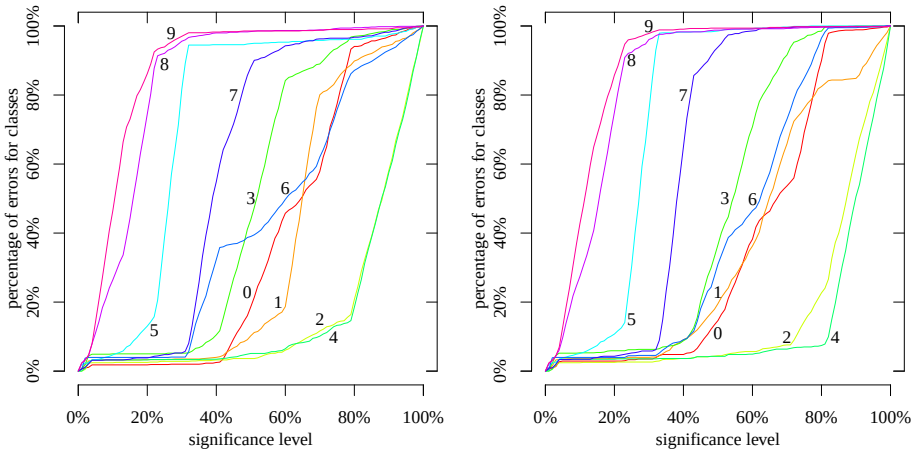

Figure 7: The final percentage of errors for different classes for (unconditional) hypergraphical conformal predictors; different colours correspond to different classes. The predictions are not class-wise valid. The left panel is for the pure exchangeability conformal predictor and the right panel for the pure hypergraphical conformal predictor. The results are for the LED data set of $10,000$ observations with 1% of noise.

the *pure hypergraphical label conditional conformal predictor* (the nontrivial hypergraphical model is used in (20) and (6)); the first two are the same predictors that we already studied in previous sections, and the last two are new.

All four predictors make predictions in the online mode, and the final percentage of errors (21) is calculated for these predictions and all significance levels between 0% and 100%. The percentage of errors plotted against the significance level will be called the *calibration graph*. For valid predictions the calibration graph is the diagonal extending from the bottom left corner (no errors at the significance level 0%, which is achieved when all prediction sets are the whole label set) to the top right corner (errors at each prediction step at the significance level 100%, which is the result of all predictions being the empty predictions).

Figure 7 shows the calibration graphs for the two unconditional conformal predictors. In each plot, the ten calibration graphs of different colours correspond to labels $\{0, 1, \dots, 9\}$. As expected, these unconditional conformal predictions are not class-wise valid. In these plots, calibration graphs that are below the diagonal correspond to easy labels (there are fewer errors than expected), and calibration graphs above the diagonal are for difficult labels (the number of errors is greater than expected). The most difficult digits are 8 and 9: this is not surprising since each of them has 3 other digits at a Hamming distance of 1 from it, more than any other digit (see Figure 2; 0, 6, and 9 are at a Hamming distance of 1 from 8, and 3, 5, and 8 are at a Hamming distance of 1 from 9). The easiest digits are 2 and 4, because these are the only digits that do not have any other digits at a Hamming distance of 1 from them (and both have 2 digits

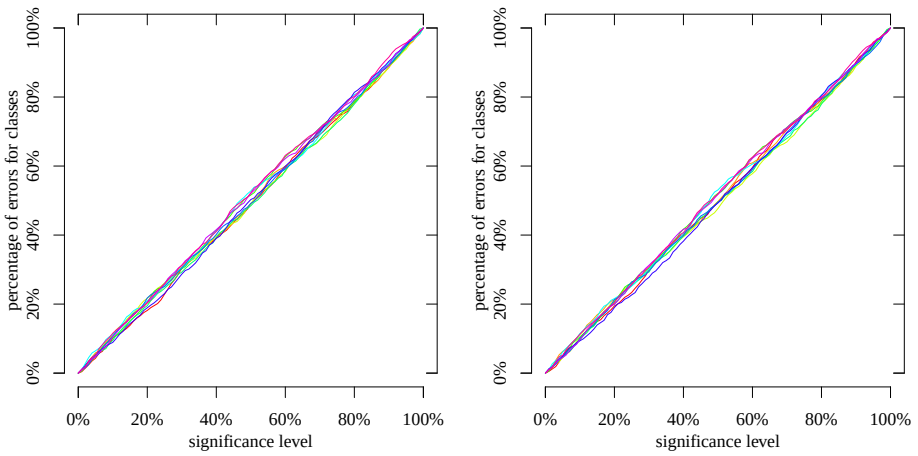

Figure 8: The same as Figure 7 for hypergraphical label conditional conformal predictors. The predictions are class-wise valid.

at a Hamming distance of 2: 3 and 8 for 2, and 1 and 9 for 4).

Figure 8 shows the results for the two label conditional conformal predictors under hypergraphical models. These predictors are constructed in order to produce class-wise valid predictions, and the experiments just confirm this property.

Efficiency

Let us compare the efficiency of conformal prediction and label conditional conformal prediction under hypergraphical models. We do so by calculating cumulative observed fuzziness OF_n (defined earlier in (10)).

Figure 9 shows the cumulative observed fuzziness for online predictions. The solid lines are for unconditional predictors, whose p-values are defined by (4), and the dash-dot lines are for label conditional predictors, whose p-values are defined by (20). Again, two models are considered: the exchangeability model and the nontrivial hypergraphical model; each model can be used for calculating the hypergraphical CP conformity scores (6) or for p-values ((4) and (20)). These combinations give four unconditional conformal predictors and four label conditional conformal predictors. Also, as before, we consider two idealized predictors: the unconditional idealized predictor is obtained using the true distribution for the data instead of the backward kernel P_{σ^*} in both (4) and (6), and analogously the *label conditional idealized predictor* is obtained using the true distribution in both (20) and (6). The notation in the legend is similar to that in the previous set of experiments (see Figures 3–6) except that "lc" stands for "label conditional".

As expected, the price to pay for the class-wise validity of label conditional conformal predictors is that they are less efficient than the corresponding uncon-

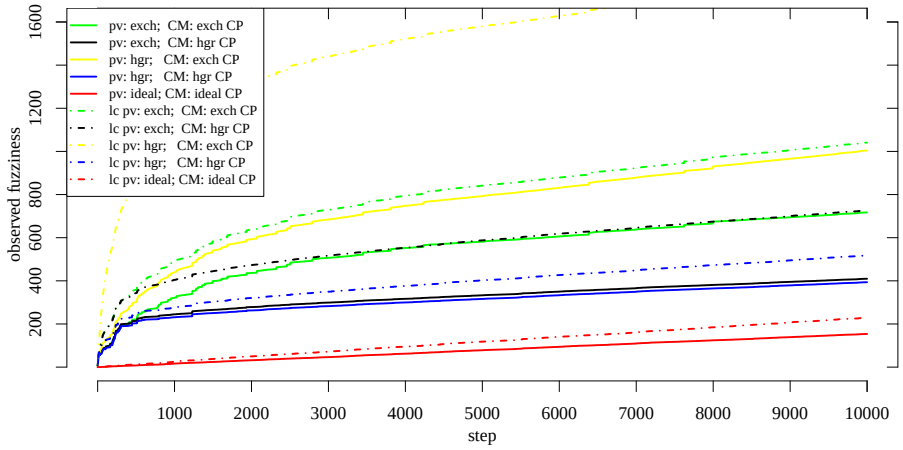

Figure 9: Observed fuzziness for unconditional conformal predictors and label conditional (lc) conformal predictors. The results are for the LED data set of $10,000$ observations with 1% of noise.

Table 5: The final values of the observed fuzziness in Figure 9 for the black and blue label conditional graphs.

Seed (10^4)	0	1	...	99	Average	St. dev.
lc pv: exch; CM: hgr CP	726.1	729.7	...	709.6	724.3	31.97
lc pv: hgr; CM: hgr CP	517.5	522.1	...	503.2	512.2	30.96

ditional conformal predictors. But the performance of the pure hypergraphical label conditional conformal predictor (the second lowest dash-dot line) is almost as good as that for the corresponding unconditional one (the second lowest solid line). The performance of the other label conditional predictors (unfortunately, including the predictor corresponding to the black line, which we recommended in the unconditional setting of the previous section) is noticeably worse than that of the corresponding unconditional ones. The final values for the black and blue graphs in Figure 9 are given in Table 5.

From the computational point of view, the label conditional way of computing p-values (20) is cheaper: for the label conditional p-values one only needs to look at the conformity scores of configurations with the same label.

6 Conclusion

The main finding of this paper is that nontrivial hypergraphical models can be useful for conformal prediction when they are true. More surprisingly, in

our experiments and the unconditional setting they only need to be used as soft models; the performance does not suffer much if the exchangeability model continues to be used as the hard model. This interesting phenomenon deserves a further empirical study.

The empirical study of label conditional conformal predictors under hypergraphical models has demonstrated that they are essential for the class-wise validity. Finally, we have seen that the performance of label conditional conformal predictors is close to that of unconditional ones if the hypergraphical models are used as both the hard and soft models.

Directions of further research include extending our approach to the regression setting (where $\mathbf{Y}$ is the set of real numbers) and the multilabel setting (where each object can belong to multiple classes).

Acknowledgements

We thank the COPA 2013 reviewers for comments that improved the results of the paper (in particular, for suggesting the CP conformity measure). We are indebted to Royal Holloway, University of London, for continued support and funding. This work has also been supported by: the EraSysBio+ grant SHIPREC from the European Union, BBSRC and BMBF; a VLA grant on machine learning algorithms; a grant from the National Natural Science Foundation of China (No. 61128003); a grant from the Cyprus Research Promotion Foundation (research contract TPE/ORIZO/0609(BIE)/24); grant EP/K033344/1 from EPSRC; a grant from AFOSR ("Semantic completions").

References

[1] K. Bache and M. Lichman. UCI machine learning repository. School of Information and Computer Sciences, University of California, Irvine, CA, USA, 2013.

[2] Vineeth N. Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk, editors. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations, and Applications*. Elsevier, Waltham, MA, 2014.

[3] Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, New York, 1999. Reprinted in 2007.

[4] Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Conformal prediction under hypergraphical models. In Harris Papadopoulos, Andreas S. Andreou, Lazaros Iliadis, and Ilias Maglogiannis, editors, *Artificial Intelligence Applications and Innovations. Second Workshop on Conformal Prediction and Its Applications*, pages 371–383, Heidelberg, 2013. Springer.

[5] Valentina Fedorova, Ilia Nouretdinov, and Alex Gammerman. Testing the Gauss linear assumption for on-line predictions. *Progress in Artificial Intelligence*, 1:205–213, 2012.

[6] Ulf Johansson, Rikard Konig, Tuve Lofstrom, and Henrik Bostrom. Evolved decision trees as conformal predictors. In Luis Gerardo de la Fraga, editor, *Proceedings of the 2013 IEEE Conference on Evolutionary Computation*, volume 1, pages 1794–1801, Cancun, Mexico, 2013.

[7] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

[8] Vladimir Vovk, Valentina Fedorova, Alex Gammerman, and Ilia Nouretdinov. Criteria of efficiency for conformal prediction, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 11, April 2014.

[9] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[10] Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. On-line predictive linear regression. *Annals of Statistics*, 37:1566–1590, 2009.

[11] Vladimir Vovk, Ivan Petej, and Valentina Fedorova. From conformal to probabilistic prediction, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 12, May 2014.