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 July 25, 2013.

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 ([6], Chapter 2). However, the definition of conformal predictors can be easily extended to a wide class of statistical models, called online compression models (OCMs; [6], 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., [6] and [2].

This paper studies conformal prediction under the OCMs known as hypergraphical models ([6], 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 of many real-world problems: for example, different symptoms can 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 [6], Chapter 8, and [4]). Hypergraphical OCMs have been used only in the context of Venn rather than conformal prediction (see [6], 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 hypergraphical 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 concludes.

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 $\mathbf{X} \times \mathbf{Y}$ are called *examples*. A *training set* is a sequence of examples $(z_1, \ldots, z_l)$, where each example $z_i = (x_i, y_i)$ 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 set. We focus on the case where $\mathbf{X}$ and $\mathbf{Y}$ are finite.

2.1 Hypergraphical Structures

In this paper we assume that examples are structured, consisting of variables. 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 set of all configurations on E. A *table*[2] on a set E is an assignment of natural numbers 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*.

2.2 Hypergraphical Online Compression Models

The example 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 at the beginning of this section.

An example $z \in \mathbf{Z}$ *agrees* with a configuration on a set $E \subseteq V$ (or the configuration agrees with the example) if the restriction $z|_E$ of z to the variables

[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.

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

in E coincides with the configuration. A table set σ *generated* by a sequence of examples $(z_1, \dots, z_n)$ assigns to each configuration on each cluster the number of examples in the sequence that agree with the configuration; the size of each table in σ will be equal to the number of examples in the sequence, and this number is called the *size* of the table set. Different sequences of examples can generate the same table set σ, and we denote $\#\sigma$ the number of different sequences generating σ.

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 example 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 $F(\sigma \downarrow z, z) = \sigma$.

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 examples $z_1, z_2, \dots$ 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 example 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 has 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. 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 ([6], 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 set $(z_1, \ldots, z_l)$ and an HOCM $(\Sigma, \Box, \mathbf{Z}, F, B)$. 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 example z w.r. to a summary σ. Intuitively, the score reflects how typical it is to observe z having the summary σ.

For each $y \in \mathbf{Y}$ denote $\sigma^* \in \Sigma$ the table set generated by the sequence $(z_1, \ldots, z_l, (x, y))$ (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 $\alpha_z := A(\sigma^* \downarrow z, z)$ (notice that $\alpha_{(x,y)}$ is always defined). The *p-value* for y, denoted $p^{(y)}$, is defined by

$$p^{(y)} := \sum_{z : \alpha_z < \alpha_{(x,y)}} P_{\sigma^*}(z) + \theta \cdot \sum_{z : \alpha_z = \alpha_{(x,y)}} P_{\sigma^*}(z) \tag{3}$$

(cf. (8.4) in [6]), where $\theta \sim \mathbf{U}[0, 1]$ is a random number 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 *conformal predictor* Γ based on A outputs the prediction set

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

(Such randomized conformal predictors were referred to as "smoothed" in [6].)

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 3.2 will describe the criteria used in this paper.

3.1 Conformity Measures for HOCM

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

$$A(\sigma, (x, y)) := P_{\sigma^*}(y \mid x) := \frac{P_{\sigma^*}\left((x, y)\right)}{\sum_{y' \in \mathbf{Y}} P_{\sigma^*}\left((x, y')\right)}, \tag{4}$$

where $\sigma^* := F(\sigma, (x, y))$ and $P_{\sigma^*}\left((x, y)\right)$ 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

$$f(x) := \max_{y \in \mathbf{Y}} P_{\sigma^*}(y \mid x), \tag{5}$$

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

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

The function maps each object x to one of the labels at which the maximum in (5) is attained. The *signed predictability conformity measure* is defined by

$$A(\sigma, (x, y)) := \begin{cases} f(x) & \text{if } y = \hat{y}(x) \\ -f(x) & \text{otherwise.} \end{cases} \tag{6}$$

3.2 Criteria for the Quality of Conformal Prediction

In this paper we study the performance of conformal predictors in the online prediction protocol (Protocol 1). Reality generates examples (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 $\Gamma_n^\epsilon := \Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$ at each significance level ϵ.

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

Protocol 1 Online prediction protocol

```
for $n = 1, 2, \dots$ do
    Reality outputs $x_n \in \mathbf{X}$
    Predictor outputs $\Gamma_n^\epsilon \subseteq \mathbf{Y}$ for all $\epsilon \in (0, 1)$
    Reality outputs $y_n \in \mathbf{Y}$
end for
```

y_n is not in Γ_n^ϵ. The validity of conformal predictors means that, for any significance level ϵ, the probability of error $y_n \notin \Gamma_n^\epsilon$ is equal to ϵ. It has been proved that conformal predictors are automatically valid under their models ([6], 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_n^ϵ over the first n steps defined by

$$\mathrm{mult}_n^\epsilon := \begin{cases} 1 & \text{if } |\Gamma_n^\epsilon| > 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \mathrm{Mult}_n^\epsilon := \sum_{i=1}^n \mathrm{mult}_i^\epsilon$$

at each significance level $\epsilon \in (0, 1)$ (cf. [6], Chapter 3). Another way is to report the *cumulative size of the prediction sets*

$$\mathrm{Size}_n^\epsilon := \sum_{i=1}^n |\Gamma_i^\epsilon|$$

at each significance level $\epsilon \in (0, 1)$. We will also consider two ways to measure the efficiency of conformal predictors that do not depend on the significance level. Let $p_n^{(y)}$, $y \in \mathbf{Y}$, be the p-values (3) used by the conformal predictor for computing the prediction set Γ_n^ϵ at the nth step of the online prediction protocol. The *cumulative unconfidence* Unconf_n over the first n steps is defined by

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

the *unconfidence* unconf_n at step n can be equivalently defined as the second largest p-value among $p_n^{(y)}$, $y \in \mathbf{Y}$. (Unconfidence is a trivial modification of the standard notion of confidence: see [6], (3.66).) Finally, the efficiency can be measured by the *cumulative sum of p-values*

$$\mathrm{pSum}_n := \sum_{i=1}^n \sum_{y \in \mathbf{Y}} p_i^{(y)}.$$

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 and [5].

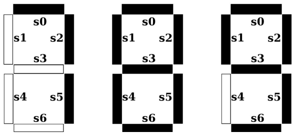

Figure 1: LED images for digits 7, 8, and 9 in the seven-segment display.

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 size of predictions* $\mathrm{Size}_n^\epsilon/n$; we would like the former to be close to 0 and the latter to be close to 1 for small significance levels.

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 Size_n^ϵ and in the sense of pSum_n.

See [5] for precise statements and proofs.

4 Experimental Results

4.1 LED Data Set

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.

Figure 1 shows several objects in the data set (these are "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 examples with noise. There is an ideal image for each digit. An example has seven binary attributes $s_0, \dots, s_6$ (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 the label to the data set.

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

$$Q\left(\{(s_0, \dots, s_6, c)\}\right) = Q_7\left(C = c\right) \cdot \prod_{i=0}^{6} Q_i\left(S_i = s_i \mid C = c\right), \qquad (7)$$

where Q_7 is the uniform distribution on the decimal digits and

$$Q_i\left(S_i = s_i \mid C = c\right) := \begin{cases} 1 - p_{\text{noise}} & \text{if } s_i = s_i^c \\ p_{\text{noise}} & \text{otherwise,} \end{cases} \qquad i = 0, \ldots, 6, \tag{8}$$

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

4.2 Hypergraphical Assumptions for LED Data Sets

We consider two hypergraphical models that agree with the decomposition (7). 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). It is clear that this is a junction tree and that $C_s = \{c\}$ for each edge s. It is also clear from (7) that the assumption (1) is satisfied; e.g., we can set

$$\begin{aligned} f_{\{s_0,c\}}\left(s_0, c\right) &:= Q_7\left(C = c\right) \cdot Q_0\left(S_0 = s_0 \mid C = c\right); \\ f_{\{s_i,c\}}\left(s_i, c\right) &:= Q_i\left(S_i = s_i \mid C = c\right), \quad i = 1, \ldots, 6. \end{aligned}$$

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 $10,000$ examples. The data are generated according to the model (7) with the probability of noise $p_{\text{noise}} = 1\%$. The data generation programs are written in C, and our data processing programs are written in R; in both cases we set the seed of the pseudorandom number generator to 0. The text below assumes that the reader can see Figures 2–5 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.

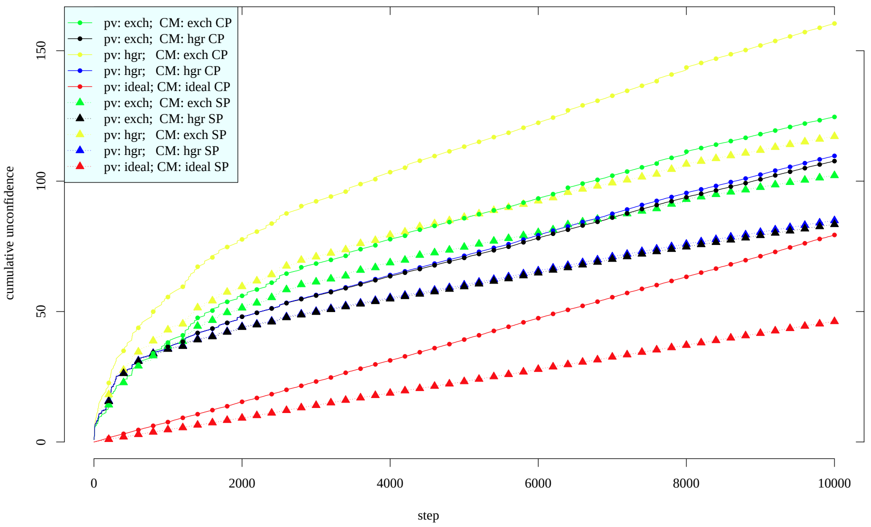

Figure 2: Cumulative unconfidence for online predictions. The results are for the LED data set with 1% of noise and 10,000 examples.

Table 1: The final values of the cumulative unconfidence in Figure 2 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

Each of the figures corresponds to an efficiency criterion for conformal predictors; namely, Figure 2 plots Unconf_n versus $n = 1, \ldots, 10000$ in the online prediction protocol, Figure 3 plots $\mathrm{pSum}_n - n/2$ versus $n = 1, \ldots, 10000$, Figure 4 plots $\mathrm{Mult}^{\epsilon}_{10000}/10000$ (the percentage of multiple predictions) versus $\epsilon \in [0, 0.05]$, and Figure 5 plots $\mathrm{Size}^{\epsilon}_{10000}/10000$ (the average size of predictions) versus $\epsilon \in [0, 0.05]$. We consider two conformity measures: the conditional probability (CP) conformity measure (4) and the signed predictability (SP) conformity measure (6). The graphs corresponding to the former are represented in our plots as lines with dots, and the graphs corresponding to the latter are represented as lines with triangles.

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

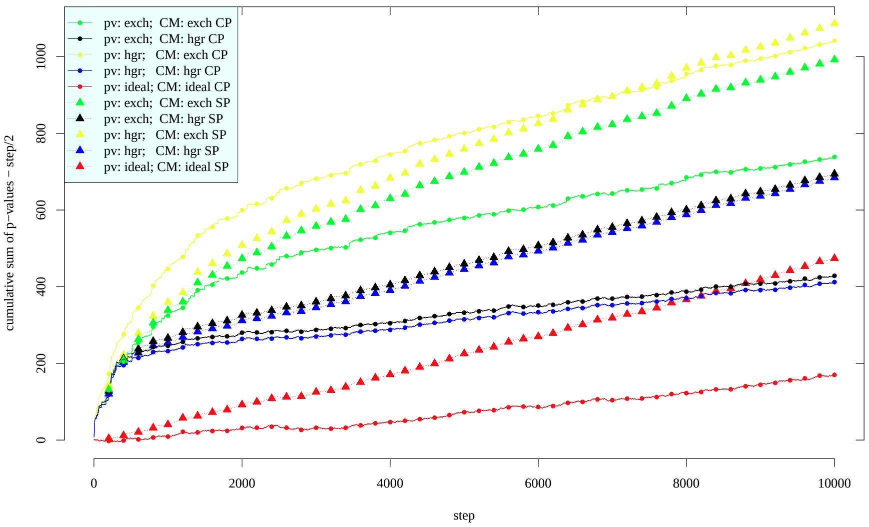

Figure 3: Adjusted cumulative sum of p-values, $\mathrm{pSum}_n - n/2$, for online predictions. The results are for the LED data set with 1% of noise and 10,000 examples.

Table 2: The final values of the adjusted cumulative sum of p-values in Figure 3 for the black and blue graphs.

Seed (10^4)	0	1	...	99	Average	St. dev.
pv: exch; CM: hgr CP	428.5	430.9	...	457.9	405.6	42.7
pv: hgr; CM: hgr CP	412.1	411.1	...	440.4	383.3	42.7
pv: exch; CM: hgr SP	694.3	755.9	...	772.4	674.8	69.9
pv: hgr; CM: hgr SP	684.7	738.7	...	756.5	656.0	69.2

drawn only for comparison, representing an unachievable ideal goal. In the idealized case we know the true distribution for the data (given by (7), (8), and $p_{\mathrm{noise}} = 1\%$). The true distribution is used instead of the backward kernel P_{σ^*} in both (3) and (4) for the CP conformity measure and in both (3) and (6) 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*

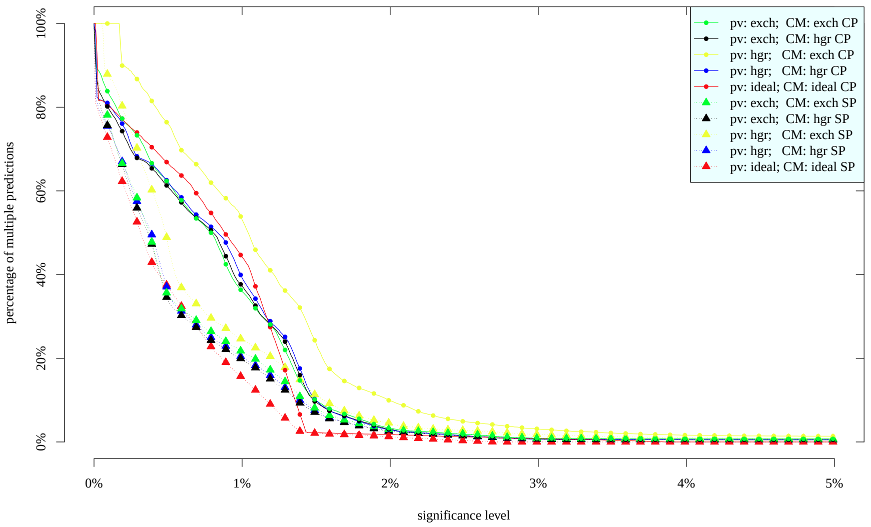

Figure 4: The final percentage of multiple predictions for significance levels between 0% and 5%. The results are for the LED data set with 1% of noise and $10,000$ examples.

Table 3: The final percentage of multiple predictions in Figure 4 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

hypergraphical conformal predictor (represented by blue lines in our plots) is obtained using the nontrivial hypergraphical model both when computing p-values (see (3)) and when computing the conformity measure ((4) in the case of CP and (6) 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 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 in [7]. The model used when computing the p-values (see (3)) is the hard model; the valid-

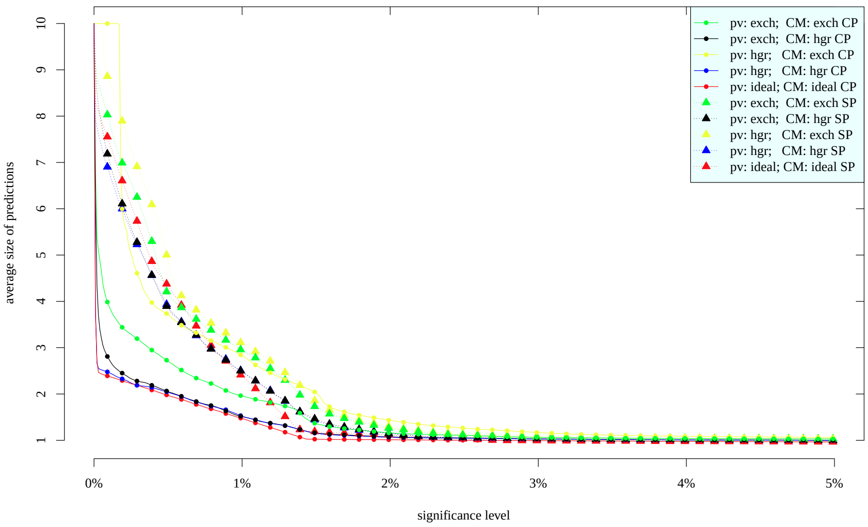

Figure 5: The final average size of predictions for significance levels between 0% and 5%. The results are for the LED data set with 1% of noise and 10,000 examples.

Table 4: The final average size of 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	1.512	1.501	...	1.509	1.535	0.124
pv: hgr; CM: hgr CP	1.520	1.477	...	1.492	1.513	0.122
pv: exch; CM: hgr SP	2.478	2.693	...	2.655	2.405	0.360
pv: hgr; CM: hgr SP	2.487	2.623	...	2.626	2.371	0.344

ity of the conformal predictor depends on it. The model used when computing conformity scores (see (4) and (6)) is the soft model; when it is violated, validity is not affected, although efficiency can suffer. The true probability distribution (7) 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 [6], Proposition 9.2.)

In the legends of Figures 2–5, 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, and "hgr" refers to the nontrivial hypergraphical model.

The most interesting graphs in Figures 2–5 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 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 2 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 2 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 3 shows the adjusted cumulative sum of p-values $\mathrm{pSum}_n - n/2$. We subtract $n/2$ since even for the best predictors the cumulative sum of p-values is at least $n/2$, up to statistical fluctuations: indeed, summing only the p-values for the true labels would already give $n/2$ (up to statistical fluctuations). For this criterion the predictors based on the CP conformity measure outperform the predictors based on the SP conformity measure (the lines with dots are below the lines of the same colour with triangles), 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 4 shows the percentage of multiple predictions after observing $10,000$ examples 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 5 shows the average size of predictions after observing $10,000$ examples 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 (or even better than) the blue CP graph, corresponding to the pure hypergraphical predictor, except for very low significance levels when the average size exceeds 2. The closeness at the significance level 1% is confirmed by Table 4.

5 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 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.

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.

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, 2013. To appear.

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

[5] Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alex Gammerman. Optimality criteria for conformal prediction. Manuscript, 2013.

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

[7] Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. On-line predictive linear regression, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 1, May 2005.