Well-calibrated predictions from on-line compression models

Vladimir Vovk

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

On-line Compression Modelling Project

Working Paper #8

April 15, 2003
revised March 19, 2004

Project web site:
http://vovk.net/kp

Abstract

It has been shown recently that Transductive Confidence Machine (TCM) is automatically well-calibrated when used in the on-line mode and provided that the data sequence is generated by an exchangeable distribution. In this paper we strengthen this result by relaxing the assumption of exchangeability of the data-generating distribution to the much weaker assumption that the data agrees with a given "on-line compression model".

Contents

1 Introduction

Transductive Confidence Machine (TCM) was introduced in (Saunders et al., 1999; Vovk et al., 1999) as a practically meaningful way of providing information about reliability of the predictions made. In (Vovk, 2002) it was shown that TCM's confidence information is valid in a strong non-asymptotic sense under the standard assumption that the examples are exchangeable. In §2 we define a general class of models, called "on-line compression models", which include not only the exchangeability model but also the Gaussian model, the Markov model, and many other interesting models. An on-line compression model (OCM) is an automaton (usually infinite) for summarizing the information about observed data efficiently. It is usually impossible to restore the data from OCM's summary (so OCM performs lossy compression), but it can be argued that the only information lost is noise, since one of our requirements is that the summary should be a "sufficient statistic". In §3 we construct "confidence transducers" and state the main result of the paper (proved in Appendix A) showing that the confidence information provided by confidence transducers is valid in a strong sense. In the last three sections, §4–6, we consider several interesting examples of on-line compression models: Gaussian, Gauss linear, Markov, exchangeability, and hypergraphical models; two of these models (Gauss linear and Markov) do not assume the exchangeability of examples. The idea of compression modelling was the main element of Kolmogorov's programme for applications of probability, which is discussed in Appendix B.

2 On-line compression models

We are interested in making predictions about a sequence of examples $z_1, z_2, \dots$ output by Nature. Typically we will want to say something about example z_n, $n = 1, 2, \dots$, given the previous examples $z_1, \dots, z_{n-1}$. In this section we will discuss an assumption that we might be willing to make about the examples, and in the next section the actual prediction algorithms.

An *on-line compression model* is a 5-tuple $M = (\Sigma, \square, \mathbf{Z}, (F_n), (B_n))$, where:

1. Σ is a measurable space called the *summary space*; its elements are called *summaries*; $\square \in \Sigma$ is a summary called the *empty summary*;

2. $\mathbf{Z}$ (the *example space*) is a measurable space from which the examples z_i are drawn;

3. F_n, $n = 1, 2, \ldots$, are functions of the type $\Sigma \times \mathbf{Z} \to \Sigma$ called *forward functions*;

4. B_n, $n = 1, 2, \ldots$, are kernels of the type $\Sigma \to \Sigma \times \mathbf{Z}$ called *backward kernels*; in other words, each B_n is a function $B_n(A \,|\, \sigma)$ which depends on $\sigma \in \Sigma$ and a measurable set $A \subseteq \Sigma \times \mathbf{Z}$ such that

   - for each σ, $B_n(A \,|\, \sigma)$ as a function of A is a probability distribution in $\Sigma \times \mathbf{Z}$;
   - for each A, $B_n(A \,|\, \sigma)$ is a measurable function of σ;

   it is required that B_n be a reverse to F_n in the sense that

   $$B_n\left(F_n^{-1}(\sigma) \,|\, \sigma\right) = 1$$

   for each $\sigma \in F_n(\Sigma \times \mathbf{Z})$. We will sometimes write $B_n(\sigma)$ for the probability distribution $A \mapsto B_n(A \,|\, \sigma)$.

Next we explain briefly the intuition behind this formal definition and introduce some further notation.

An OCM is a way of summarizing statistical information. At the beginning we do not have any information, which is represented by the empty summary $\sigma_0 := \square$. When the first example z_1 arrives, we update our summary to $\sigma_1 := F_1(\sigma_0, z_1)$, etc.; when example z_n arrives, we update the summary to $\sigma_n := F_n(\sigma_{n-1}, z_n)$. This process is represented in Figure 1. Let t_n be the *nth statistic* in the OCM, which maps the sequence of the first n examples $z_1, \ldots, z_n$ to σ_n:

$$\begin{aligned} &t_1(z_1) := F_1(\sigma_0, z_1); \\ &t_n(z_1, \ldots, z_n) := F_n(t_{n-1}(z_1, \ldots, z_{n-1}), z_n), \quad n = 2, 3, \ldots \ . \end{aligned}$$

The value $t_n(z_1, \ldots, z_n)$ is a summary of the full data sequence $z_1, \ldots, z_n$ available at the end of trial n; our definition requires that the summaries should be computable on-line: the function F_n updates σ_{n-1} to σ_n.

Condition 3 in the definition of OCM reflects its on-line character, as explained in the previous paragraph. We want, however, the system of summarizing statistical information represented by the OCM to be efficient, so

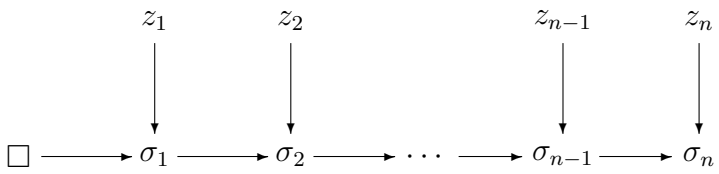

Figure 1: Using the forward functions F_n to compute σ_n from $z_1, \ldots, z_n$

that no useful information is lost. This is reflected in Condition 4: the distribution B_n of the more detailed description (σ_{n-1}, z_n) given the less detailed σ_n is known and so does not carry any information about the distribution generating the examples $z_1, z_2, \ldots$; in other words, σ_n contains the same useful information as (σ_{n-1}, z_n), and the extra information in (σ_{n-1}, z_n) is noise. This intuition would be captured in statistical terminology (see, e.g., Cox and Hinkley 1974, §2.2) by saying that σ_n is a "sufficient statistic" of $z_1, \ldots, z_n$ (although this expression does not have a formal meaning in our present context, since we do not have a full statistical model $\{P_\theta : \theta \in \Theta\}$).

Analogously to Figure 1, we can compute the distribution of the data sequence $z_1, \ldots, z_n$ from σ_n (see Figure 2). Formally, using the kernels $B_n(d\sigma_{n-1}, dz_n \,|\, \sigma_n)$, we can define the "conditional distribution" P_n of $z_1, \ldots, z_n$ given σ_n by the formula

$$P_n(A_1 \times \cdots \times A_n \,|\, \sigma_n) := \int \cdots \int B_1(A_1 \,|\, \sigma_1) B_2(d\sigma_1, A_2 \,|\, \sigma_2) \ldots B_{n-1}(d\sigma_{n-2}, A_{n-1} \,|\, \sigma_{n-1}) B_n(d\sigma_{n-1}, A_n \,|\, \sigma_n) \quad (1)$$

for each product set $A_1 \times \cdots \times A_n$, $A_i \subseteq \mathbf{Z}$, $i = 1, \ldots, n$. (We will use the expression "conditional distribution" for P_n despite the fact that in general it is *not* obtained from some other probability distribution by conditioning.)

We say that a probability distribution P in $\mathbf{Z}^\infty$ *agrees* with the OCM $(\Sigma, \square, \mathbf{Z}, (F_n), (B_n))$ if, for each n, $B_n(A \,|\, \sigma)$ is a version of the conditional probability, w.r. to P, that $(t_{n-1}(z_1, \ldots, z_{n-1}), z_n) \in A$ given $t_n(z_1, \ldots, z_n) = \sigma$ and given the values of $z_{n+1}, z_{n+2}, \ldots$.

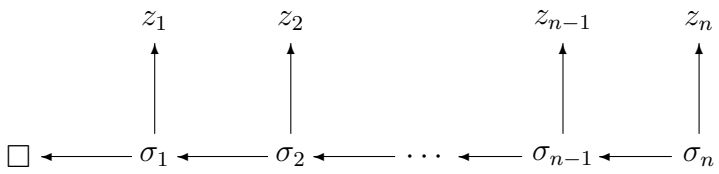

Figure 2: Using the backward functions B_n to extract the distribution of $z_1, \dots, z_n$ from σ_n

3 Confidence transducers and the main result

A *randomised transducer* is a function f of the type $(\mathbf{Z} \times [0,1])^* \to [0,1]$. It is called "transducer" because it can be regarded as mapping each input sequence $(z_1, \theta_1, z_2, \theta_2, \dots)$ in $(\mathbf{Z} \times [0,1])^\infty$ (the examples z_i are complemented by random numbers θ_i) into the output sequence $(p_1, p_2, \dots)$ defined by $p_n := f(z_1, \theta_1, \dots, z_n, \theta_n)$, $n = 1, 2, \dots$; we will say that $p_1, p_2, \dots$ are the *p-values* produced by the randomised transducer. We say that the randomised transducer f is *valid* w.r. to an OCM M if the output p-values $p_1 p_2 \dots$ are always distributed according to the uniform distribution U^∞ in $[0,1]^\infty$, provided the input examples $z_1 z_2 \dots$ are generated by a probability distribution that agrees with M and $\theta_1 \theta_2 \dots$ are generated, independently of $z_1 z_2 \dots$, from U^∞. If we drop the dependence on the random numbers θ_n, we obtain the notion of *deterministic transducer*.

Any sequence of measurable functions $A_n : \Sigma \times \mathbf{Z} \to \mathbb{R}$, $n = 1, 2, \dots$, is called an *individual strangeness measure* w.r. to the OCM $M = (\Sigma, \square, \mathbf{Z}, (F_n), (B_n))$. The *confidence transducer* associated with (A_n) is the deterministic transducer where p_n are defined as

$$p_n := B_n \left(\{ (\sigma, z) \in \Sigma \times \mathbf{Z} : A_n(\sigma, z) \geq A_n(\sigma_{n-1}, z_n) \} \mid \sigma_n \right) \tag{2}$$

and

$$\sigma_n := t_n(z_1, \dots, z_n), \quad \sigma_{n-1} := t_{n-1}(z_1, \dots, z_{n-1}).$$

The randomised version is obtained by replacing (2) with

$$p_n := B_n\left(\{(\sigma, z) \in \Sigma \times \mathbf{Z} : A_n(\sigma, z) > A_n(\sigma_{n-1}, z_n)\} \,|\, \sigma_n\right) + \theta_n B_n\left(\{(\sigma, z) \in \Sigma \times \mathbf{Z} : A_n(\sigma, z) = A_n(\sigma_{n-1}, z_n)\} \,|\, \sigma_n\right). \quad (3)$$

A *confidence transducer in* an OCM M is a confidence transducer associated with some individual strangeness measure w.r. to M.

Theorem 1 *Suppose the examples $z_n \in \mathbf{Z}$, $n = 1, 2, \ldots$, are generated from a probability distribution P that agrees with an on-line compression model. Any randomised confidence transducer in that model is valid (will produce independent p-values p_n distributed uniformly in $[0, 1]$).*

Confidence transducers can be used for "prediction with confidence". Suppose each example z_n consists of two components, x_n (the object) and y_n (the label); at trial n we are given x_n and the goal is to predict y_n. Therefore, $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$, where $\mathbf{X}$ is the *object space* and $\mathbf{Y}$ is the *label space*.

One mode of prediction with confidence is "region prediction" (as in Vovk 2002). Suppose we are given a *significance level* $\delta > 0$ (the maximum probability of error we are prepared to tolerate). When given x_n, we can output as the *prediction region* $\Gamma_n^{(\delta)} \subseteq \mathbf{Y}$ the set of labels y such that $y_n = y$ would lead to a p-value $p_n > \delta$: e.g., in the randomised case,

$$\Gamma_n^{(\delta)} := \{y \in \mathbf{Y} : f(z_1, \theta_1, \ldots, z_{n-1}, \theta_{n-1}, (x_n, y), \theta_n) > \delta\},$$

where f is the randomised transducer being used and $\theta_1, \theta_2, \ldots$ are the random numbers. When a confidence transducer is applied in this mode, it is referred to as a *Transductive Confidence Machine*. If error at trial n is defined as $y_n \notin \Gamma_n^{(\delta)}$, then by Theorem 1 errors at different trials are independent and the probability of error at each trial is δ, assuming the p_n are produced by a randomised confidence transducer. In particular, such region predictors are *well-calibrated*, in the sense that the number E_n of errors made in the first n trials satisfies

$$\lim_{n\to\infty} \frac{E_n}{n} = \delta.$$

This implies that if the p_n are produced by a deterministic confidence transducer, we will still have the conservative version of this property,

$$\overline{\lim_{n\to\infty}} \frac{E_n}{n} \le \delta.$$

An alternative way of presenting the confidence transducer's output (used in Vovk et al. 1999 and several other papers) is reporting, after seeing x_n, a *predicted label* $\hat{y}_n \in \arg\max_{y \in \mathbf{Y}} p_n(y)$, the *confidence* $1 - p_n^{(2)}$, and the *credibility* $p_n^{(1)}$, where $p_n(y)$ is the p-value that would be obtained if $y_n = y$, $p_n^{(1)}$ is the largest value among $p_n(y)$, and $p_n^{(2)}$ is the second largest value among $p_n(y)$.

4 Gaussian model

The only special case of OCM studied from the point of view of prediction with confidence before this paper was the exchangeability model; this model, together with its powerful generalization that we call the "hypergraphical model", will be discussed in §6. We start with two new models, Gaussian (this section) and Markov (§5). Many more models are considered in (Bernardo and Smith, 2000, Chapter 4). For defining specific OCM, we will specify their statistics t_n and conditional distributions P_n; these will uniquely identify F_n and B_n.

In the Gaussian model, $\mathbf{Z} := \mathbb{R}$, the statistics are

$$t_n(z_1, \dots, z_n) := (\overline{z}_n, R_n)\,, \tag{4}$$

where

$$\overline{z}_n := \frac{1}{n} \sum_{i=1}^{n} z_i, \quad R_n := \sqrt{(z_1 - \overline{z}_n)^2 + \dots + (z_n - \overline{z}_n)^2},$$

and $P_n(dz_1, \dots, dz_n \,|\, \sigma)$ is the uniform distribution in $t_n^{-1}(\sigma)$ (in other words, for $\sigma = (\overline{z}_n, R_n)$, it is the uniform distribution in the $(n-2)$-dimensional sphere in $\mathbb{R}^n$ with centre $(\overline{z}_n, \dots, \overline{z}_n) \in \mathbb{R}^n$ of radius R_n lying inside the hyperplane $\frac{1}{n}(z_1 + \dots + z_n) = \overline{z}_n$).

It is clear that there are many possible representations of essentially the same model; for example, we obtain an equivalent model if we replace (4) by

$$t_n(z_1, \dots, z_n) := \left(\sum_{i=1}^{n} z_i, \sum_{i=1}^{n} z_i^2 \right). \tag{5}$$

Let us give an explicit expression of the prediction region for the Gaussian model and individual strangeness measure

$$A_n(\sigma_{n-1}, z_n) = A_n((\overline{z}_{n-1}, R_{n-1}), z_n) := |z_n - \overline{z}_{n-1}| \tag{6}$$

(it is easy to see that this individual strangeness measure is equivalent, in the sense of leading to the same p-values, to $|z_n - \overline{z}_n|$, as well as to several other natural expressions, including (7)). Under $P_n(dz_1, \ldots, dz_n \,|\, \sigma)$ and assuming $n > 2$, the expression

$$\sqrt{\frac{(n-1)(n-2)}{n}} \frac{z_n - \overline{z}_{n-1}}{R_{n-1}} \tag{7}$$

has Student's t-distribution with $n-2$ degrees of freedom. (This fact is proven in, e.g., Cramér 1946, §29.4, where it is assumed, however, that $z_1, \ldots, z_n$ are independent and have the same Gaussian distribution. The latter assumption is easy to replace by our assumption of the uniform distribution; for a general argument, see the proof of Proposition 1 below.) Let $t_{\epsilon,k}$ be the value defined by $\mathbb{P}\{\tau > t_{\epsilon,k}\} = \epsilon$, where τ has Student's t-distribution with k degrees of freedom. The prediction region (or *prediction interval*, in this case) corresponding to the individual strangeness measure (6) and a significance level δ is the set of z satisfying

$$|z - \overline{z}_{n-1}| \le t_{\delta/2,n-2} \sqrt{\frac{n}{(n-1)(n-2)}} R_{n-1}. \tag{8}$$

Therefore, we obtained the usual prediction regions based on the t-test (as in Baker 1935, Wilks 1941, and, implicitly, Fisher 1925); now, however, we can see that the errors of this standard procedure (applied in the on-line fashion) are independent.

Gauss linear model

We will now consider a rich extension of the Gaussian model. In the *Gauss linear model*, the example space is of the "regression type", $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ with the label space being the real line, $\mathbf{Y} := \mathbb{R}$, and the object space being the p-dimensional Euclidean space, $\mathbf{X} := \mathbb{R}^p$. The statistics are

$$t_n(x_1, y_1, \ldots, x_n, y_n) := \left(x_1, \ldots, x_n, \sum_{i=1}^n y_i x_i, \sum_{i=1}^n y_i^2 \right) \tag{9}$$

(so Σ can be set to $\mathbf{X}^* \times \mathbb{R}^p \times \mathbb{R}$), and each conditional distribution $P_n(\cdot \,|\, \sigma)$ is the uniform probability distribution in the sphere $t_n^{-1}(\sigma)$ (we consider a point to be a sphere; typically $t_n^{-1}(\sigma_n)$ will be a point unless $n > p$).

The Gaussian model in the form (5) is a special case (using, however, a different notation, z_i for y_i) corresponding to $p = 1$ and x_n restricted to $x_n = 1$, $n = 1, 2, \ldots$. Using $\sum_{i=1}^n y_i x_i$ rather than $\sum_{i=1}^n y_i$ reflects the possibility that y_i can depend on x_i.

It is clear that the probability distribution for $z_1, z_2, \ldots$ in the linear regression statistical model

$$y_n = w \cdot x_n + \xi_n, \tag{10}$$

where $w \in \mathbb{R}^p$ is a constant vector and ξ_n are independent variables with the same zero-mean Gaussian distribution $\mathcal{N}(0, \sigma^2)$, always agrees with the Gauss linear model. The name "Gauss linear model" was suggested (in a similar context) by Seal (1967).

Our next proposition will use the following notation: $\hat{y}_i^n$ is the least-squares prediction for the object x_i based on the examples $z_1, \ldots, z_n$; $\hat{y}_n$ is a shorthand for $\hat{y}_n^{n-1}$; X_l, $l = 1, 2, \ldots$, is the $l \times p$ matrix whose ith row is x_i' (i.e., x_i transposed), $i = 1, \ldots, l$; and

$$\hat{\sigma}_l^2 := \frac{1}{l-p} \sum_{i=1}^{l} (y_i - \hat{y}_i^l)^2$$

is the standard estimate of σ^2 from the first l examples.

Proposition 1 *The prediction region based on the nonconformity measure $A_n := |y_n - \hat{y}_n|$ is given, for $n > p + 1$ satisfying $\operatorname{rank}(X_{n-1}) = p$, by the formula*

$$\Gamma_n^{(\delta)} = \left[\hat{y}_n - t_{\delta/2, n-p-1} V_n, \hat{y}_n + t_{\delta/2, n-p-1} V_n \right], \tag{11}$$

where

$$V_n := \sqrt{\hat{\sigma}_{n-1}^2 (1 + x_n' (X_{n-1}' X_{n-1})^{-1} x_n)}.$$

Proof It is a standard fact (see, e.g., Stuart et al. 1999, §32.10) that $(y_n - \hat{y}_n)/V_n$ has the t-distribution with $n - p - 1$ degrees of freedom; this assumes, however, the standard model (10) rather than the uniform conditional distribution of the Gauss linear model. Let us check that $(y_n - \hat{y}_n)/V_n$ will still have the t-distribution with $n - p - 1$ degrees of freedom under the uniform conditional distribution.

First note that $(y_n - \hat{y}_n)/V_n$ can be rewritten so that it only depends on the *n-residuals* $y_i - \hat{y}_i^n$ (i.e., residuals computed from all n examples $z_1, \ldots, z_n$).

Indeed, a standard statistical result (Montgomery et al., 2001, (4.12)) shows that

$$\hat{\sigma}_{n-1}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i^n)^2 - (y_n - \hat{y}_n^n)^2/(1 - x_n'(X_n'X_n)^{-1}x_n)}{n-p-1}; \tag{12}$$

another standard result (Montgomery et al., 2001, (4.11)) shows that

$$y_n - \hat{y}_n = \frac{y_n - \hat{y}_n^n}{1 - x_n'(X_n'X_n)^{-1}x_n}. \tag{13}$$

Let $Y_n := (y_1, \dots, y_n)'$ be the vector of the first n labels and $\hat{Y}_n := (\hat{y}_1, \dots, \hat{y}_n)'$ be the vector of the first n least-squares predictions. According to the geometric interpretation of the least squares method in the standard model (10) (see, e.g., Draper and Smith 1998, Chapters 20–21), the vector of n-residuals is distributed symmetrically around $\hat{Y}_n$ in the space orthogonal to the estimation space $\{X_n w : w \in \mathbb{R}^p\}$. On the other hand, according to (9) and the definition of P_n, $P_n(\cdot \mid \sigma_n)$ is the uniform distribution on the sphere, of radius equal to the length of the vector of n-residuals, in the hyperplane orthogonal to the estimation space and passing through the projection $\hat{Y}_n$ of Y_n onto the estimation space. Since the ratio $(y_n - \hat{y}_n)/V_n$ (expressed through the n-residuals $y_i - \hat{y}_i^n$) does not change if all n-residuals are multiplied by the same positive constant (and, therefore, its distribution does not change if the random vector of n-residuals is scaled to have a given length), we may replace the Gaussian distribution of (10) by our uniform distribution $P_n(\cdot \mid \sigma_n)$.

The proof will be complete if we show that

$$\left| \frac{y_n - \hat{y}_n}{V_n} \right| = \frac{|y_n - \hat{y}_n|}{V_n}$$

is a bona fide individual strangeness measure which monotonically increases as $|y_n - \hat{y}_n|$ increases for any fixed $\sigma_n := t_n(z_1, \dots, z_n)$. This is simple: standard statistical formulas show that $|y_n - \hat{y}_n|/V_n$ is expressible through $t_{n-1}(x_1, y_1, \dots, x_{n-1}, y_{n-1})$ and $z_n = (x_n, y_n)$, and, from (12) and (13),

$$\frac{|y_n - \hat{y}_n|}{V_n} \uparrow\uparrow \frac{|y_n - \hat{y}_n^n|}{\sqrt{C - c(y_n - \hat{y}_n^n)^2}} \uparrow\uparrow \frac{(y_n - \hat{y}_n^n)^2}{C - c(y_n - \hat{y}_n^n)^2}$$

$$\uparrow\downarrow \frac{C - c(y_n - \hat{y}_n^n)^2}{(y_n - \hat{y}_n^n)^2} \uparrow\uparrow \frac{1}{(y_n - \hat{y}_n^n)^2} \uparrow\downarrow |y_n - \hat{y}_n^n| \uparrow\uparrow |y_n - \hat{y}_n|,$$

where $C > 0$ and c are constants (for a fixed σ_n), $\uparrow\uparrow$ means "changes in the same direction", and $\uparrow\downarrow$ means "changes in the opposite direction". ∎

It is easy to check that Proposition 1 contains (8) as a special case.

The prediction interval (11) is standard (see, e.g., Montgomery et al. 2001, (3.54)), but Theorem 1 adds the usual extra feature: the independence of errors in the on-line setting.

Remark The methods of this subsection are applicable to time series, although only to the simplest ones: e.g., if

$$y_n = f(n) + \cos \frac{n-a}{T} + \xi_n$$

where $f(n)$ is a polynomial of a known order p, T is a known constant (the period of the seasonal component) and ξ_n are independent and identically distributed zero-mean Gaussian random variables, we can set

$$x_n := \left(1, n, \dots, n^p, \cos \frac{n}{T}, \sin \frac{n}{T}\right)$$

and use formula (11). Constructing good TCM in more interesting cases would require new methods.

5 Markov model

In this section we assume that the example space $\mathbf{Z}$ is finite. The following notation for digraphs will be used: $\mathrm{in}(v)$ (resp. $\mathrm{out}(v)$) stands for the number of arcs entering (resp. leaving) vertex v; $n_{u,v}$ is the number of arcs leading from vertex u to vertex v.

The *Markov summary* of a data sequence $z_1 \dots z_n$ is the following digraph with two vertices marked:

- the set of vertices is $\mathbf{Z}$;
- the vertex z_1 is marked as the *source* and the vertex z_n is marked as the *sink* (these two vertices are not necessarily distinct);
- the arcs of the digraph are the transitions $z_i z_{i+1}$, $i = 1, \dots, n-1$; the arc $z_i z_{i+1}$ has z_i as its tail and z_{i+1} as its head.

It is clear that in any such digraph all vertices v satisfy $\mathrm{in}(v) = \mathrm{out}(v)$ with the possible exception of the source and sink (unless they coincide), for which we then have $\mathrm{out}(\mathrm{source}) = \mathrm{in}(\mathrm{source}) + 1$ and $\mathrm{in}(\mathrm{sink}) = \mathrm{out}(\mathrm{sink}) + 1$. We

will call a digraph with this property a *Markov graph* if the arcs with the same tail and head are indistinguishable (for example, we do not distinguish two Eulerian paths that only differ in the order in which two such arcs are passed); its underlying digraph will have the same structure but all its arcs will be considered to have their own identity. Markov summaries will always be regarded as Markov graphs.

The Markov model is the OCM with the nth statistic $\sigma_n = t_n(z_1, \dots, z_n)$ equal to the Markov summary of $z_1, \dots, z_n$ and the conditional probability distribution $P_n(\cdot \mid \sigma_n)$ being the uniform distribution over the Eulerian paths in the Markov graph σ_n (with each Eulerian path represented by the sequence of vertices along it).

This is the explicit definition of the Markov model as an OCM $(\Sigma, \square, \mathbf{Z}, F, B)$:

- $\mathbf{Z}$ is a finite set; its elements (examples) are also called *states*;
- $\Sigma \setminus \{\square\}$ is the set of all Markov graphs with the vertex set $\mathbf{Z}$;
- $\square$ is, e.g., the empty set;
- $F_n(\sigma, z)$, $n = 2, 3, \dots$, is the Markov graph obtained from σ by adding an arc from σ's sink to z and making z the new sink; $F_1(\square, z)$ is the Markov graph with no arcs and with both source and sink at z;
- let $\sigma \downarrow z$, where σ is a Markov graph and z is one of σ's vertices, be the Markov graph obtained from σ by removing an arc from z to σ's sink ($\sigma \downarrow z$ does not exist if there is no arc from z to σ's sink) and moving the sink to z, and let $N(\sigma)$ be the number of Eulerian paths from the source to the sink in a Markov graph σ; $B_n(\sigma)$ is $(\sigma \downarrow z, \text{sink})$ with probability $N(\sigma \downarrow z)/N(\sigma)$, where sink is σ's sink and z ranges over the states for which $\sigma \downarrow z$ is defined.

Notice that any Markov probability distribution in $\mathbf{Z}^\infty$ (i.e., a probability distribution P such that, for some function $g : \mathbf{Z}^2 \to [0, 1]$, the conditional probability that $z_n = z$ given $z_1, \dots, z_{n-1}$ always equals $g(z_{n-1}, z)$) agrees with the Markov model.

We will take as the individual strangeness measure

$$A_n(\sigma, z) := -B_n\left(\{(\sigma, z)\} \mid F_n(\sigma, z)\right) \tag{14}$$

(we need the minus sign because lower probability makes an example stranger). To give a computationally efficient representation of the confidence transducer corresponding to this individual strangeness measure, we need the following two graph-theoretic results, versions of the BEST theorem and the Matrix-Tree theorem, respectively.

Lemma 1 *In any Markov graph $\sigma = (V, E)$ the number of Eulerian paths from the source to the sink equals*

$$T(\sigma)\frac{\operatorname{out}(\text{sink})\prod_{v\in V}(\operatorname{out}(v)-1)!}{\prod_{u,v\in V} n_{u,v}!},$$

where $T(\sigma)$ is the number of spanning out-trees in the underlying digraph rooted at the source.

Lemma 2 *To find the number $T(\sigma)$ of spanning out-trees rooted at the source in the underlying digraph of a Markov graph σ with vertices $z_1, \dots, z_n$ (z_1 being the source),*

- *create the $n \times n$ matrix with the elements $a_{i,j} = -n_{z_i,z_j}$;*
- *change the diagonal elements so that each column sums to 0;*
- *compute the co-factor of $a_{1,1}$.*

These two lemmas immediately follow from Theorems VI.24 and VI.28 in (Tutte, 2001).

It is now easy to obtain an explicit formula for prediction in the binary case $\mathbf{Z} = \{0, 1\}$. First we notice that, for $n > 1$,

$$B_n(\{(\sigma \downarrow z, \text{sink})\} \mid \sigma) = \frac{N(\sigma \downarrow z)}{N(\sigma)} = \frac{T(\sigma \downarrow z) n_{z,\text{sink}}}{T(\sigma)\operatorname{out}(\text{sink})}$$

(all $n_{u,v}$ refer to the numbers of arcs in σ and sink is σ's sink; we set $N(\sigma \downarrow z) = T(\sigma \downarrow z) := 0$ when $\sigma \downarrow z$ does not exist). The following simple corollary from the last formula is sufficient for computing the probabilities B_n in the binary case:

$$B_n(\{(\sigma \downarrow \text{sink}, \text{sink})\} \mid \sigma) = \frac{n_{\text{sink},\text{sink}}}{\operatorname{out}(\text{sink})}.$$

This gives us the following formulas for the TCM in the binary Markov model (remember that the individual strangeness measure is (14)). Suppose

the current summary is given by a Markov graph with $n_{i,j}$ arcs going from vertex i to vertex j ($i, j \in \{0, 1\}$) and let $f : [0, 1] \to [0, 1]$ be the function that squashes $[0.5, 1]$ to 1:

$$f(p) := \begin{cases} p & \text{if } p < 0.5 \\ 1 & \text{otherwise.} \end{cases}$$

If the current sink is 0, the p-value corresponding to the next example 0 is

$$f\left(\frac{n_{0,0} + 1}{n_{0,0} + n_{0,1} + 1}\right)$$

and the p-value corresponding to the next example 1 is (with $0/0 := 1$)

$$f\left(\frac{n_{1,0}}{n_{1,0} + n_{1,1}}\right). \tag{15}$$

If the current sink is 1, the p-value corresponding to the next example 1 is

$$f\left(\frac{n_{1,1} + 1}{n_{1,1} + n_{1,0} + 1}\right)$$

and the p-value corresponding to the next example 0 is (with $0/0 := 1$)

$$f\left(\frac{n_{0,1}}{n_{0,1} + n_{0,0}}\right).$$

Figure 3 shows the result of a computer simulation; as expected, the error line is close to the straight line with the slope close to the significance level.

6 Exchangeability and hypergraphical models

The exchangeability model has statistics

$$t_n(z_1, \dots, z_n) := \wr z_1, \dots, z_n \int;$$

given the value of the statistic, all orderings have the same probability $1/n!$. Formally, the set of bags $\wr z_1, \dots, z_n \int$ of size n is defined as $\mathbf{Z}^n$ equipped with the σ-algebra of symmetric (i.e., invariant under permutations of components) events; the distribution on the orderings is given by $z_{\pi(1)}, \dots, z_{\pi(n)}$, where $z_1, \dots, z_n$ is a fixed ordering and π is a random permutation (each permutation is chosen with probability $1/n!$).

The main results of (Vovk, 2002) and (Vovk et al., 2003) are special cases of Theorem 1.

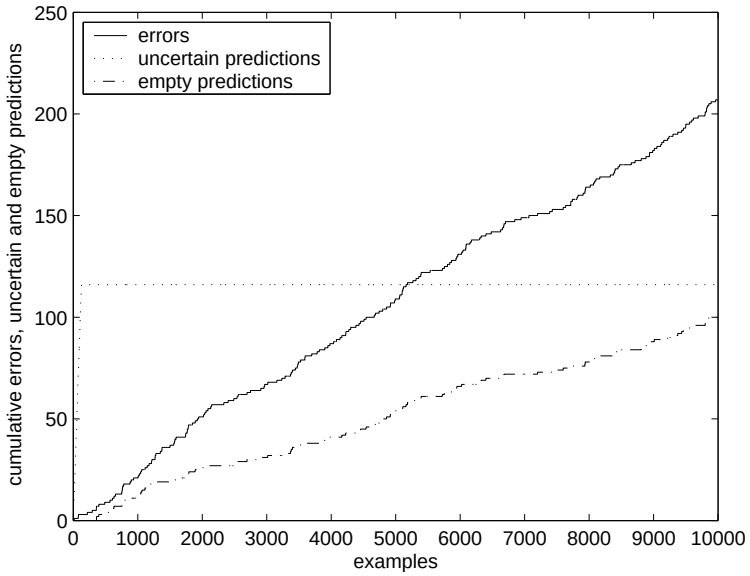

Figure 3: TCM predicting the binary Markov chain with transition probabilities $\mathbb{P}(1\,|\,0) = \mathbb{P}(0\,|\,1) = 1\%$ at significance level 2%; the cumulative numbers of errors (prediction regions not covering the true label), uncertain (i.e., containing more than one label) and empty prediction regions are shown

Hypergraphical structures

We now assume that the examples are structured, consisting of "variables". Formally, a *hypergraphical structure* is a triple $(V, \mathcal{E}, \Xi)$, where:

- V is a finite set whose elements will be called *variables*;
- $\mathcal{E}$ is a family of V's subsets; elements of $\mathcal{E}$ are called *clusters*; the union of all clusters is required to be the whole of V;
- Ξ is a function that maps each variable $v \in V$ into the finite set $\Xi(v)$ of the "values that v can take"; $\Xi(v)$ is called the *frame* of v; to exclude trivial cases, we always assume $\forall v : |\Xi(v)| > 1$.

In applications one (or more) of the variables is marked as the *label*, but this will not be used in our considerations. A *configuration* on a cluster (or, more generally, V's subset) E is an assignment of an element of $\Xi(v)$ to each

$v \in E$. An *example* is a configuration on V; we take $\mathbf{Z}$ to be the set of all examples.

A *table* on a cluster E is an assignment of a non-negative number to each configuration on E. We will mainly be interested in *natural tables*, which assign only natural (i.e., non-negative integer) numbers to configurations. (These are known as "contingency tables" in statistics.) The *size* of the table is the sum of values that it assigns to different configurations. A *table set* assigns to each cluster a table on that cluster. We will only be interested in table sets all of whose tables have the same size, which is then called the size of the table set. The number assigned by a table set σ to a configuration of a cluster E will sometimes be called the σ-*count* of that configuration.

Hypergraphical models

Now we are ready to define the OCM associated with a hypergraphical structure $(V, \mathcal{E}, \Xi)$; as usual, the most intuitive definition is in terms of statistics t_n and conditional distributions P_n. The table set $t_n(z_1, \ldots, z_n)$ *generated* by a data sequence $z_1, \ldots, z_n$ assigns to each configuration the number of examples among $z_1, \ldots, z_n$ that *agree* with that configuration (we say that an example z *agrees* with a configuration on a cluster E if that configuration coincides with the restriction $z|_E$ of z to E). The number of data sequences generating a table set σ will be denoted $N(\sigma)$ (for $N(\sigma)$ to be non-zero the size of σ must exist, and then the length of each sequence generating σ will be equal to its size). The table sets σ with $N(\sigma) > 0$ (called *consistent* table sets) are called *summaries*; they form the summary space Σ of the hypergraphical on-line compression model associated with $(V, \mathcal{E}, \Xi)$. The conditional probability distribution $P_n(\cdot \,|\, \sigma)$, where n is the size of σ, is the uniform distribution in the set of all data sequences $z_1, \ldots, z_n$ that generate σ.

The explicit definition of the hypergraphical model $(\Sigma, \square, \mathbf{Z}, F, B)$ is as follows:

- Σ is the set of all summaries (i.e., consistent table sets); $\square$ is the *empty* table set, i.e., the one of size 0;
- $\mathbf{Z}$ is the set of all examples (i.e., configurations on V);
- the table set $F(\sigma, z)$ is obtained from σ by adding 1 to the σ-count of each configuration consistent with z;

- an example z is *consistent* with a summary σ if the σ count of each configuration that agrees with z is positive; if so, we define $\sigma \downarrow z$ from σ by subtracting 1 from the σ-count of any configuration that agrees with z; $B_n(\sigma)$, where n is the size of σ, is defined by

$$B(\{(\sigma \downarrow z, z)\} \,|\, \sigma) := \frac{N(\sigma \downarrow z)}{N(\sigma)}.$$

Among the probability distributions P that agree with the hypergraphical model with structure $(V, \mathcal{E}, \Xi)$ are power distributions p^∞ such that each p (a probability distribution in $\mathbf{Z}$) decomposes into

$$p\{z : z(v) = a(v), \forall v \in V\} = \prod_{E \in \mathcal{E}} f_E(z|_E), \tag{16}$$

where a is any configuration on V, f is a fixed table set, and $z|_E$ is, as usual, the restriction of z to E.

The exchangeability model with the example space $\mathbf{Z}$ corresponds to the hypergraphical model with only one cluster, $\mathcal{E} = \{V\}$.

Junction-tree models

An important special case is where we can arrange the clusters of a hypergraphical structure into a "junction tree". We will be able to give relatively efficient prediction algorithms only for such junction-tree structures. Fortunately, modelling with junction-tree structures is a well-developed field; for example, the standard way of dealing with Bayesian networks is to transform them into junction-tree structures (see, e.g., Jensen 1996).

Formally, a *junction tree* for a hypergraphical structure $(V, \mathcal{E}, \Xi)$ is an undirected tree (U, S) (with U the set of vertices and S the set of edges) together with a bijective mapping C from the vertices U of the tree to the clusters $\mathcal{E}$ of the hypergraphical structure which satisfies the following property: if a vertex v lies on the path from a vertex u to a vertex w in the tree (U, S), then

$$C_u \cap C_w \subseteq C_v$$

(we let C_x stand for $C(x)$). The tree (U, S) will also sometimes be called the junction tree (when the bijection is clear from the context). It is convenient to identify vertices v of the junction tree with the corresponding clusters C_v

in $\mathcal{E}$. If $s = \{u, v\} \in S$ is an edge of the junction tree connecting vertices u and v, we will write C_s for $C_u \cap C_v$; C_s will be called the *separator* between C_u and C_v.

We will say "junction-tree structures/models" to mean hypergraphical structures/models in which the clusters are arranged into a junction tree.

It is easy to characterize consistent table sets in junction-tree structures. If $E_1 \subseteq E_2 \subseteq V$ and f is a table set on E_2, its *marginalisation* to E_1 is the table f^* on E_1 such that $f^*(a) = \sum_b f(b)$ for all configurations a on E_1, where b ranges over all configurations on E_2 consistent with a (i.e., such that $b|_{E_1} = a$).

Lemma 3 *Let $(V, \mathcal{E}, \Xi)$ be a junction-tree structure. A natural table set σ on $(V, \mathcal{E}, \Xi)$ is consistent if and only if the following two conditions hold:*

- *each table in σ is of the same size;*
- *if clusters $E_1, E_2 \in \mathcal{E}$ intersect, the marginalisations of their tables to $E_1 \cap E_2$ coincide.*

This lemma is obvious; it, however, ceases to be true if the assumption that $(V, \mathcal{E}, \Xi)$ is a junction-tree structure is dropped. (Indeed, suppose V consists of three binary variables A, B, C, $\mathcal{E}$ consists of the clusters AB, AC, and BC, and consider the table set assigning 1 to the configurations $A = 0\,\&\,B = 0$, $A = 1\&B = 1$, $A = 0\&C = 0$, $A = 1\&C = 1$, $B = 0\&C = 0$, $B = 1\&C = 1$, and assigning 0 to all other configurations. The two conditions hold but the table is not consistent.)

If σ is a summary and E is a cluster, we let σ_E stand for the table that σ assigns to E. If E is a separator, say $E = C_{\{u,v\}}$, σ_E stands for the marginalisation of σ_{C_u} (equivalently, by Lemma 3, of σ_{C_v}) to E.

The *factorial-product* of a cluster or separator E in a summary σ is, by definition,

$$\mathrm{fp}_\sigma(E) := \prod_{a \in \mathrm{conf}(E)} \sigma_E(a)!,$$

where $\mathrm{conf}(E)$ is the set of all configurations on E.

Lemma 4 *Consider a summary σ of size n on a junction-tree structure. The number of data sequences of length n compatible with the table set σ equals*

$$N(\sigma) = \frac{n! \prod_{s \in S} \mathrm{fp}_\sigma(C_s)}{\prod_{u \in U} \mathrm{fp}_\sigma(C_u)}. \tag{17}$$

Proof The proof is by induction in the size of the junction tree. If the junction tree consists of only one vertex u, the formula (17) becomes

$$\frac{n!}{\mathrm{fp}_\sigma(C_u)} = \frac{n!}{\prod_{a\in\mathrm{conf}(C_u)} \sigma_{C_u}(a)!},$$

which is the correct multinomial coefficient.

Now let us assume that (17) is true for some tree and prove that it remains true for that tree extended by adding an edge s and a vertex u. (The example space for the new tree will be bigger.) We are required to show that the number of data sequences consistent with σ is multiplied by

$$\frac{\mathrm{fp}_\sigma(C_s)}{\mathrm{fp}_\sigma(C_u)} = \prod_{a\in\mathrm{conf}(C_s)} \frac{\sigma_{C_s}(a)!}{\prod_{b\in\mathrm{comp}(a)} \sigma_{C_u}(b)!}, \tag{18}$$

where $\mathrm{comp}(a)$ is the set of all configurations on C_u compatible with a. It remains to notice that the number of ways in which each sequence of n examples in the old tree can be extended to a sequence of n examples in the new tree is given by the right-hand side of (18). ∎

Lemma 5 *Given the summary σ of the first n examples, the $B_n(\sigma)$-probability that $z_n = a$ equals*

$$\frac{\prod_{u\in U} \sigma_{C_u}(a|_{C_u})}{n \prod_{s\in S} \sigma_{C_s}(a|_{C_s})} \tag{19}$$

(this ratio is set to 0 if any of the factors in the numerator or denominator is 0; in this case $z_n = a$ is incompatible with the summary σ).

Proof Letting fp' stand for the factorial-product in the summary $\sigma \downarrow a$, we obtain for the probability of $z_n = a$:

$$\frac{N(\sigma \downarrow a)}{N(\sigma)} = \frac{(n-1)! \prod_{s\in S} \mathrm{fp}'_\sigma(C_s) \prod_{u\in U} \mathrm{fp}_\sigma(C_u)}{\prod_{u\in U} \mathrm{fp}'_\sigma(C_u) n! \prod_{s\in S} \mathrm{fp}_\sigma(C_s)} = \frac{\prod_{u\in U} \sigma_{C_u}(a|_{C_u})}{n \prod_{s\in S} \sigma_{C_s}(a|_{C_s}).}$$ ∎

The reader may recognize (19) as the maximum likelihood estimate of p under (16). This simple representation of $B_n(\sigma)$ makes it possible to compute p-values (which can then be used for prediction with confidence) using Monte Carlo simulation. Another powerful technique that can be applied to sampling from $B_n(\sigma)$ is described in (Diaconis and Sturmfels, 1998).

Acknowledgments

I am grateful to Satoshi Aoki, Phil Dawid, Alex Gammerman, Per Martin-Löf, Glenn Shafer, Akimichi Takemura, participants in the workshop "Statistical Learning in Classification and Model Selection" (January 2003, Eurandom), and ALT'2003 participants for useful discussions. The anonymous referees' comments about the conference version of this paper helped to improve the presentation. Gregory Gutin's advice about graph theory is gratefully appreciated.

This work was partially supported by EPSRC (grant GR/R46670/01), BBSRC and EU (grant IST-1999-10226).

References

Asarin, E. A. (1987), Some properties of Kolmogorov δ-random finite sequences, *Theory of Probability and its Applications* **32**, 507–508.

Asarin, E. A. (1988), On some properties of finite objects random in the algorithmic sense, *Soviet Mathematics Doklady* **36**, 109–112.

Baker, G. A. (1935), The probability that the mean of a second sample will differ from the mean of a first sample by less than a certain multiple of the standard deviation of the first sample, *Annals of Mathematical Statistics* **6**, 197–201.

Bernardo, J. M. and Smith, A. F. M. (2000), *Bayesian Theory*, Wiley, Chichester.

Cox, D. R. and Hinkley, D. V. (1974), *Theoretical Statistics*, Chapman and Hall, London.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ.

Diaconis, P. and Sturmfels, B. (1998), Algebraic algorithms for sampling from conditional distributions, *Annals of Statistics* **26**, 363–397.

Draper, N. R. and Smith, H. (1998), *Applied Regression Analysis*, third edn, Wiley, New York.

Fisher, R. A. (1925), Applications of "Student's" distribution, *Metron* **5**, 90–104.

Freedman, D. A. (1962), Invariants under mixing which generalise de Finetti's theorem, *Annals of Mathematical Statistics* **33**, 916–923.

Jensen, F. V. (1996), *An Introduction to Bayesian Networks*, UCL Press, London.

Kolmogorov, A. N. (1968), Logical basis for information theory and probability theory, *IEEE Transactions of Information Theory* **IT-14**, 662–664.

Kolmogorov, A. N. (1983), Combinatorial foundations of information theory and the calculus of probabilities, *Russian Mathematical Surveys* **38**, 29–40.

Lauritzen, S. L. (1988), *Extremal Families and Systems of Sufficient Statistics*, Vol. 49 of *Lecture Notes in Statistics*, Springer, New York.

Martin-Löf, P. (1966), The definition of random sequences, *Information and Control* **9**, 602–619.

Martin-Löf, P. (1974), Repetitive structures and the relation between canonical and microcanonical distributions in statistics and statistical mechanics, *in* O. Barndorff-Nielsen, P. Blæsild and G. Schou, eds, 'Proceedings of Conference on Foundational Questions in Statistical Inference', Aarhus, pp. 271–294.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2001), *Introduction to Linear Regression Analysis*, third edn, Wiley, New York.

Nouretdinov, I., V'yugin, V. and Gammerman, A. (2003), Transductive Confidence Machine is universal, *in* R. Gavaldà, K. P. Jantke and E. Takimoto, eds, 'Proceedings of the Fourteenth International Conference on Algorithmic Learning Theory', Vol. 2842 of *Lecture Notes in Artificial Intelligence*, Springer, Berlin.

Saunders, C., Gammerman, A. and Vovk, V. (1999), Transduction with confidence and credibility, *in* T. Dean, ed., 'Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence', Vol. 2, Morgan Kaufmann, pp. 722–726.

Seal, H. L. (1967), Studies in the history of probability and statistics. XV: The historical development of the Gauss linear model, *Biometrika* **54**, 1–24.

Shiryaev, A. N. (1996), *Probability*, second edn, Springer, New York.

Stuart, A., Ord, K. J. and Arnold, S. (1999), *Kendall's Advanced Theory of Statistics*, Vol. 2a: Classical inference and the linear model, sixth edn, Arnold, London.

Tutte, W. T. (2001), *Graph Theory*, Cambridge University Press, Cambridge, UK.

Vovk, V. (1986), On the concept of the Bernoulli property, *Russian Mathematical Surveys* **41**, 247–248.

Vovk, V., Gammerman, A. and Saunders, C. (1999), Machine-learning applications of algorithmic randomness, *in* 'Proceedings of the Sixteenth International Conference on Machine Learning', Morgan Kaufmann, San Francisco, CA, pp. 444–453.

Vovk, V. and Shafer, G. (2003), Kolmogorov's contributions to the foundations of probability, *Problems of Information Transmission* **39**, 21–31.

Vovk, V. (2002), On-line Confidence Machines are well-calibrated, On-line Compression Modelling project, `http://vovk.net/kp`, Working Paper #1.

Vovk, V. (2002), Asymptotic optimality of Transductive Confidence Machine, On-line Compression Modelling project, `http://vovk.net/kp`, Working Paper #2.

Vovk, V. (2002), Universal well-calibrated algorithm for on-line classification, On-line Compression Modelling project, `http://vovk.net/kp`, Working Paper #3.

Vovk, V., Nouretdinov, I. and Gammerman, A. (2003), Testing exchangeability on-line, On-line Compression Modelling project, `http://vovk.net/kp`, Working Paper #5.

Wilks, S. S. (1941), Determination of sample sizes for setting tolerance limits, *Annals of Mathematical Statistics* **12**, 91–96.

A Appendix: Proof of Theorem 1

We will use the notation $\mathbb{E}_{\mathcal{F}}$ for the conditional expectation w.r. to a σ-algebra $\mathcal{F}$; if necessary, the underlying probability distribution will be given as an upper index. Similarly, $\mathbb{P}_{\mathcal{F}}$ will stand for the conditional probability w.r. to $\mathcal{F}$. In this appendix we will use the following properties of conditional expectation (see, e.g., Shiryaev 1996, §II.7.4):

A. If $\mathcal{G}$ and $\mathcal{F}$ are σ-algebras, $\mathcal{G} \subseteq \mathcal{F}$, ξ and η are bounded $\mathcal{F}$-measurable random variables, and η is $\mathcal{G}$-measurable, $\mathbb{E}_{\mathcal{G}}(\xi\eta) = \eta\,\mathbb{E}_{\mathcal{G}}(\xi)$ a.s.

B. If $\mathcal{G}$ and $\mathcal{F}$ are σ-algebras, $\mathcal{G} \subseteq \mathcal{F}$, and ξ is a random variable, $\mathbb{E}_{\mathcal{G}}(\mathbb{E}_{\mathcal{F}}(\xi)) = \mathbb{E}_{\mathcal{G}}(\xi)$ a.s.; in particular, $\mathbb{E}(\mathbb{E}_{\mathcal{F}}(\xi)) = \mathbb{E}(\xi)$.

Proof of the Theorem

This proof is a generalization of the proof of Theorem 1 in (Vovk, 2002), with the same basic idea: to show that $(p_1, \dots, p_N)$ is distributed as U^N (it is easy to get rid of the assumption of a fixed horizon N), we reverse the time. Let P be the distribution generating the examples; it is assumed to agree with the OCM. Imagine that the sample $(z_1, \dots, z_N)$ is generated in two steps: first, the summary σ_N is generated from some probability distribution (namely, the image of the distribution P generating $z_1, z_2, \dots$ under the mapping t_N), and then the sample $(z_1, \dots, z_N)$ is chosen randomly from $P_N(\cdot \mid \sigma_N)$. Already the second step ensures that, conditionally on knowing σ_N (and, therefore, unconditionally), the sequence $(p_N, \dots, p_1)$ is distributed as U^N. Indeed, roughly speaking (i.e., ignoring borderline effects), p_N will be the p-value corresponding to the statistic A_N and so distributed, at least approximately, as U (see, e.g., Cox and Hinkley 1974, §3.2); when the pair (σ_{N-1}, z_N) is disclosed, the value p_N will be settled; conditionally on knowing σ_{N-1} and z_N, p_{N-1} will also be distributed as U, and so on.

We start the formal proof by defining the σ-algebra $\mathcal{G}_n$, $n = 0, 1, 2, \dots$, as the one on the sample space $(\mathbf{Z} \times [0,1])^\infty$ generated by the random elements $\sigma_n, z_{n+1}, \theta_{n+1}, z_{n+2}, \theta_{n+2}, \dots$. In particular, $\mathcal{G}_0$ (the most informative σ-algebra) coincides with the original σ-algebra on $(\mathbf{Z} \times [0,1])^\infty$; $\mathcal{G}_0 \supseteq \mathcal{G}_1 \supseteq \cdots$.

Fix a randomised confidence transducer f; it will usually be left implicit in our notation. Let p_n be the random variable $f(z_1, \theta_1, \dots, z_n, \theta_n)$ for each $n = 1, 2, \dots$; $\mathbb{P}$ will refer to the probability distribution $P \times U^\infty$ (over examples

z_n and random numbers θ_n) and $\mathbb{E}$ to the expectation w.r.t. to $\mathbb{P}$. The proof will be based on the following lemma.

Lemma 6 *For any trial n and any $\delta \in [0,1]$,*

$$\mathbb{P}_{\mathcal{G}_n}\{p_n \le \delta\} = \delta. \tag{20}$$

Proof Let us fix a summary σ_n of the first n examples $(z_1, \dots, z_n) \in \mathbf{Z}^n$; we will omit the condition "$\mid \sigma_n$". For every pair $(\tilde\sigma, \tilde z)$ from $F_n^{-1}(\sigma_n)$ define

$$p^+(\tilde\sigma, \tilde z) := B_n\{(\sigma, z) : A_n(\sigma, z) \ge A_n(\tilde\sigma, \tilde z)\},$$
$$p^-(\tilde\sigma, \tilde z) := B_n\{(\sigma, z) : A_n(\sigma, z) > A_n(\tilde\sigma, \tilde z)\}.$$

It is clear that always $p^- \le p^+$. Notice that the semi-closed intervals $[p^-(\tilde\sigma, \tilde z), p^+(\tilde\sigma, \tilde z))$, $(\tilde\sigma, \tilde z) \in F_n^{-1}(\sigma_n)$, either coincide or are disjoint; it is also easy to see that they "lie next to each other", in the sense that their union is also a semi-closed interval (namely, $[0,1)$).

Let us say that a pair $(\tilde\sigma, \tilde z)$ is

- *strange* if $p^+(\tilde\sigma, \tilde z) \le \delta$
- *ordinary* if $p^-(\tilde\sigma, \tilde z) > \delta$
- *borderline* if $p^-(\tilde\sigma, \tilde z) \le \delta < p^+(\tilde\sigma, \tilde z)$.

We will use the notation $p^- := p^-(\tilde\sigma, \tilde z)$ and $p^+ := p^+(\tilde\sigma, \tilde z)$ where $(\tilde\sigma, \tilde z)$ is any borderline example. Notice that the B_n-measure of strange examples is p^-, the B_n-measure of ordinary examples is $1 - p^+$, and the B_n-measure of borderline examples is $p^+ - p^-$.

By the definition of randomised confidence transducer, $p_n \le \delta$ if the pair (σ_{n-1}, z_n) is strange, $p_n > \delta$ if the pair is ordinary, and $p_n \le \delta$ with probability

$$\frac{\delta - p^-}{p^+ - p^-}$$

if the pair is borderline; indeed, in this case

$$p_n = p^- + \theta_n(p^+ - p^-),$$

and so $p_n \le \delta$ is equivalent to

$$\theta_n \le \frac{\delta - p^-}{p^+ - p^-}.$$

Therefore, the overall probability that $p_n \le \delta$ is

$$p^- + (p^+ - p^-)\frac{\delta - p^-}{p^+ - p^-} = \delta. \qquad \blacksquare$$

The other basic result that we will need is the following lemma.

Lemma 7 *For any trial $n = 1, 2, \dots$, p_n is $\mathcal{G}_{n-1}$-measurable.*

Proof This follows from the definition, (3): p_n is defined in terms of σ_{n-1}, z_n and θ_n. The only technicality that might not be immediately obvious is that the function

$$B_n(\{A_n > c\} \,|\, \sigma)$$

of $c \in \mathbb{R}$ and $\sigma \in \Sigma$ is measurable. Let $C \in \mathbb{R}$. The set

$$\{(c, \sigma) : B_n(\{A_n > c\} \,|\, \sigma) > C\} \tag{21}$$

is measurable since it can be represented as

$$\bigcup_{d \in \mathbb{Q}} (0, d) \times \Sigma_d,$$

where $\mathbb{Q}$ is the set of rational numbers and Σ_c is the set of σ satisfying the inequality in (21). $\blacksquare$

Fix temporarily positive integer N. First we prove that, for any $n = 1, \dots, N$ and any $\delta_1, \dots, \delta_n \in [0, 1]$,

$$\mathbb{P}_{\mathcal{G}_n}\{p_n \le \delta_n, \dots, p_1 \le \delta_1\} = \delta_n \cdots \delta_1 \quad \text{a.s.} \tag{22}$$

The proof is by induction in n. For $n = 1$, (22) immediately follows from Lemma 6. For $n > 1$ we obtain, making use of Lemmas 6 and 7, properties A and B of conditional expectations, and the inductive assumption:

$$\begin{aligned}
&\mathbb{P}_{\mathcal{G}_n}\{p_n \le \delta_n, \dots, p_1 \le \delta_1\} \\
&\qquad = \mathbb{E}_{\mathcal{G}_n}\left(\mathbb{E}_{\mathcal{G}_{n-1}}\left(\mathbb{I}_{\{p_n \le \delta_n\}}\mathbb{I}_{\{p_{n-1} \le \delta_{n-1}, \dots, p_1 \le \delta_1\}}\right)\right) \\
&\qquad = \mathbb{E}_{\mathcal{G}_n}\left(\mathbb{I}_{\{p_n \le \delta_n\}}\, \mathbb{E}_{\mathcal{G}_{n-1}}\left(\mathbb{I}_{\{p_{n-1} \le \delta_{n-1}, \dots, p_1 \le \delta_1\}}\right)\right) \\
&\qquad\qquad = \mathbb{E}_{\mathcal{G}_n}\left(\mathbb{I}_{\{p_n \le \delta_n\}}\delta_{n-1} \cdots \delta_1\right) = \delta_n\delta_{n-1} \cdots \delta_1
\end{aligned}$$

($\mathbb{I}_E$ being the indicator of event E) almost surely.

By property B, (22) immediately implies

$$\mathbb{P}\{p_N \le \delta_N, \dots, p_1 \le \delta_1\} = \delta_N \cdots \delta_1.$$

Therefore, we have proved that the distribution of the random sequence $p_1 p_2 \cdots \in [0,1]^\infty$ coincides with U^∞ on the σ-algebra $\mathcal{F}_N$ generated by the first N coordinate random variables $p_1, \dots, p_N$. It is well known (see, e.g., Shiryaev 1996, Theorem II.3.3) that this implies that the distribution of $p_1 p_2 \dots$ coincides with U^∞ on all measurable sets in $[0,1]^\infty$.

B Appendix: Kolmogorov's programme and repetitive structures

In this section we briefly discuss Kolmogorov's programme for applications of probability and two related developments originated by Martin-Löf and Freedman. In particular, we formally define a version of the notion of repetitive structure which is in a sense isomorphic to our notion of OCM.

Kolmogorov's programme

The standard approach to modelling uncertainty is to choose a family of probability distributions (*statistical model*) one of which is believed to be the true distribution generating, or explaining in a satisfactory way, the data. (In some applications of probability theory, the true distribution is assumed to be known, and so the statistical model is a one-element set. In Bayesian statistics, the statistical model is complemented by another element, a prior distribution on the distributions in the model.) All modern applications of probability depend on this scheme.

In 1965–1970 Kolmogorov suggested a different approach to modelling uncertainty based on information theory; its purpose was to provide a more direct link between the theory and applications of probability. His main idea was that "practical conclusions of probability theory can be substantiated as implications of hypotheses of *limiting*, under given constraints, complexity of the phenomena under study" (Kolmogorov, 1983). The main features of Kolmogorov's programme can be described as follows:

C (Compression): One fixes a "sufficient statistic" for the data. This is a function of the data that extracts, intuitively, all useful information

from the data. This can be the number of ones in a binary sequence (the "Bernoulli model" in Kolmogorov 1968; Martin-Löf 1966), the number of ones after ones, ones after zeros, zeros after ones and zeros after zeros in a binary sequence (the "Markov model" in Kolmogorov 1983), the sample average and sample variance of a sequence of real numbers (the "Gaussian model" in Asarin 1987, 1988).

A (Algorithmic): If the value of the sufficient statistic is known, the information left in the data is noise. This is formalized in terms of Kolmogorov complexity: the complexity of the data under the constraint given by the value of the sufficient statistic should be maximal (in other words, the data should be *algorithmically random* given the value of the sufficient statistic).

U (Uniformity): Semantically, the requirement of algorithmic randomness in the previous item means that the conditional distribution of the data given the sufficient statistic is uniform.

D (Direct): It is preferable to deduce properties of data sets directly from the assumption of limiting complexity, without a detour through standard statistical models (examples of such direct inferences are given in Asarin 1987, 1988 and hinted at in Kolmogorov 1983), especially that Kolmogorov's models are not completely equivalent to standard statistical models (Vovk, 1986).

Kolmogorov's only two publications on his programme are (Kolmogorov, 1968, 1983); the work reported in (Martin-Löf, 1966; Vovk, 1986; Asarin, 1987, 1988) was done under his supervision by his PhD students.

After 1965 Kolmogorov and Martin-Löf worked on the information-theoretic approach to probability applications independently of each other, but arrived at similar concepts and definitions. Martin-Löf (1974) introduced the notion of *repetitive structure*, later studied by Lauritzen (1988). Martin-Löf's theory of repetitive structures has features C and U of Kolmogorov's programme but not features A and D. An extra feature of repetitive structures is their *on-line character*: the conditional probability distributions are required to be consistent and the sufficient statistic can usually be updated recursively as new data arrives.

The absence of algorithmic complexity and randomness from Martin-Löf's theory does not look surprising; e.g., it is argued in (Vovk and Shafer, 2003)

that these algorithmic notions are powerful sources of intuition, but for stating mathematical results in their strongest and most elegant form it is often necessary to "translate" them into a non-algorithmic form.

A more important deviation from Kolmogorov's ideas seems to be the absence of "direct inferences". The goal in the theory of repetitive structures is to derive standard statistical models from repetitive structures (in the asymptotic on-line setting the difference between Kolmogorov-type and standard models often disappears); to apply repetitive structures to reality one still needs to go through statistical models. In our approach (see Theorem 1 above or the optimality results in Vovk 2002,?) statistical models become irrelevant; in principle, all results can be stated without them.

Freedman and Diaconis independently came up with ideas similar to Kolmogorov's (Freedman's first paper in this direction was published in 1962); they were inspired by de Finetti's theorem and the Krylov–Bogolyubov approach to ergodic theory.

Kolmogorov only considered the three main models (exchangeability, Markov, Gaussian) that we discuss in §4–6, but many other models have been considered by later authors (see, e.g., Bernardo and Smith 2000, Chapter 4).

The difference between standard statistical modelling and Kolmogorov's modelling discussed in (Vovk, 1986) is not important for the purpose of one-step-ahead forecasting in the exchangeable case (in particular, for both exchangeability and Gaussian models of this paper; see Nouretdinov et al. 2003); it becomes important, however, in the Markov case. The theory of prediction with confidence has a dual goal: validity (there should not be too many errors) and efficiency (there should not be too many uncertain predictions, in the case of classification). In the asymmetric Markov case, although we have the validity result (Theorem 1), there is little hope of obtaining an optimality result analogous to those of (Vovk, 2002,?). A manifestation of the difference between the two approaches to modelling is, e.g., the fact that (15) involves the ratio $n_{1,0}/(n_{1,0}+n_{1,1})$ rather than something like $n_{0,1}/(n_{0,0}+n_{0,1})$.

Repetitive structures

Let Σ and $\mathbf{Z}$ be measurable spaces (of "summaries" and "examples", respectively). A *repetitive structure* contains the following two elements:

- a system of statistics (measurable functions) $t_n : \mathbf{Z}^n \to \Sigma$, $n = 1, 2, \dots$;

- a system of kernels P_n of the type $\Sigma \to \mathbf{Z}^n$, $n = 1, 2, \ldots$.

These two elements are required to satisfy the following consistency requirements:

Agreement between P_n and t_n: for each $\sigma \in t_n(\mathbf{Z}^n)$, the probability distribution $P_n(\cdot \,|\, \sigma)$ is concentrated on the set $t_n^{-1}(\sigma)$;

Consistency of t_n over n: for all integers $n > 1$, $t_n(z_1, \ldots, z_n)$ is determined by $t_{n-1}(z_1, \ldots, z_{n-1})$ and z_n, in the sense that the function t_n is measurable w.r. to the σ-algebra generated by t_{n-1} and z_n.

Consistency of P_n over n: for all integers $n > 1$ and all $\sigma \in t_n(\mathbf{Z}^n)$, $P_{n-1}(\cdot \,|\, \tau)$ should be a version of the conditional distribution of $z_1, \ldots, z_{n-1}$ when $z_1, \ldots, z_n$ is generated from $P_n(dz_1, \ldots, dz_n \,|\, \sigma)$ and it is known that $t_{n-1}(z_1, \ldots, z_{n-1}) = \tau$ and $z_n = z$ (τ ranging over $t_{n-1}(\mathbf{Z}^{n-1})$ and z over $\mathbf{Z}$).

The notions of OCM and repetitive structure are very close. If $M = (\Sigma, \square, \mathbf{Z}, (F_n), (B_n))$ is an OCM, then $M' := (\mathbf{Z}, \Sigma, (t_n), (P_n))$, as defined in §2, is a repetitive structure. If $M = (\mathbf{Z}, \Sigma, (t_n), (P_n))$ is a repetitive structure, an OCM $M' := (\Sigma', \square, \mathbf{Z}, (F_n), (B_n))$ can be defined as follows:

- $\square$ is, say, the empty set; $\Sigma' := \Sigma \cup \{\square\}$;
- F_n is a measurable function mapping $t_{n-1}(z_1, \ldots, z_{n-1})$ (interpreted as $\square$ for $n = 1$) and z_n to $t_n(z_1, \ldots, z_n)$, for all $(z_1, \ldots, z_n) \in \mathbf{Z}^n$ (the existence of such F_n follows from the consistency of t_n over n);
- $B_n(d\sigma_{n-1}, dz_n \,|\, \sigma_n)$ is the image of the distribution $P_n(dz_1, \ldots, dz_n \,|\, \sigma_n)$ under the mapping $(z_1, \ldots, z_n) \mapsto (\sigma_{n-1}, z_n)$, where $\sigma_{n-1} := t_{n-1}(z_1, \ldots, z_{n-1})$.

If M is a repetitive structure, M'' is essentially the same as M, and if M is an OCM, M'' is essentially the same as M (M and M'' can only differ on irrelevant parts of Σ: e.g., in how $P_n(\sigma)$ is defined for $\sigma \notin t_n(\mathbf{Z}^n)$).

In our examples (Gaussian, Markov, exchangeability models and their modifications) we found it more convenient to start from the corresponding repetitive structure (the statistics t_n and conditional distributions P_n); the conditions of consistency were obviously satisfied in those cases.

On-line Compression Modelling Project
Working Papers

1. *On-line confidence machines are well-calibrated*, by Vladimir Vovk. April 2002. [FOCS'2002]

2. *Asymptotic optimality of Transductive Confidence Machine*, by Vladimir Vovk. May 2002. [ALT'2002]

3. *Universal well-calibrated algorithm for on-line classification*, by Vladimir Vovk. November 2002. [COLT'2003]

4. *Mondrian Confidence Machine*, by Vladimir Vovk, David Lindsay, Ilia Nouretdinov and Alex Gammerman. March 2003.

5. *Testing exchangeability on-line*, by Vladimir Vovk, Ilia Nouretdinov and Alex Gammerman. February 2003. [ICML'2003]

6. *Criterion of calibration for Transductive Confidence Machine with limited feedback*, by Ilia Nouretdinov and Vladimir Vovk. April 2003. [ALT'2003]

7. *Online region prediction with real teachers*, by Daniil Ryabko, Vladimir Vovk and Alex Gammerman. March 2003.

8. *Well-calibrated predictions from on-line compression models*, by Vladimir Vovk. April 2003, revised March 2004. [ALT'2003]

9. *Self-calibrating probability forecasting*, by Vladimir Vovk, Glenn Shafer and Ilia Nouretdinov. June 2003. [NIPS'2003]

Versions of some of these working papers have been or will be published in conference proceedings (given in brackets).