Nonparametric predictive distributions based on conformal prediction

Vladimir Vovk
(joint work with Jieli Shen, Valery Manokhin, Min-ge Xie, Ilia Nouretdinov, and Alex Gammerman)

Royal Holloway, University of London
Rutgers University

COPA 2017, Stockholm, 14 June 2017

This talk

- Our statistical model (nonparametric): the observations are IID; standard in machine learning.
- Our goal: probabilistic regression (we want a predictive distribution for an unknown label $y \in \mathbb{R}$). As opposed to probabilistic classification (Paolo's tutorial on Venn prediction yesterday).
- The method of conformal prediction is applicable and guarantees validity (correct coverage probabilities).
- But we also want efficiency (high concentration of the predictive distribution).
- This talk: covers the paper in the Proceedings plus one or two further results.

My plan

The Dempster–Hill procedure

- There is a famous nonparametric procedure in classical statistics; I will call it the Dempster–Hill procedure.
- Its discoverers valued it highly but it was limited (there are only labels, no objects).
- Roughly, the procedure suggests using the empirical distribution function as the predictive distribution.

Two quotes

Bruce Hill, 1988:

> *Let me conclude by observing that $A_{(n)}$ is supported by all of the serious approaches to statistical inference. It is Bayesian, fiducial, and even a confidence/tolerance procedure. It is simple, coherent, and plausible. It can even be argued, I believe, that $A_{(n)}$, along with $H_{(n)}$, constitutes the fundamental solution to the problem of induction.*

Christian Genest and Jack Kalbfleisch, 1988 (in reply to Hill):

> *To be truly useful, however, the methods need extension to regression models with unknown regression parameters.*

The procedure

In Hill's words (1968):

> *A_n asserts that conditional upon the observations $X_1, \dots, X_n$, the next observation X_{n+1} is equally likely to fall in any of the open intervals between successive order statistics of the given sample.*

Frank Coolen: nonparametric predictive inference.

"LSPM" generalizes the Dempster–Hill procedure to regression.

The setting

- We have observations $z_i = (x_i, y_i)$ consisting of objects $x_i \in \mathbf{X}$ and their labels $y_i \in \mathbb{R}$.
- The object space $\mathbf{X}$ is a measurable space; the observation space is $\mathbf{Z} := \mathbf{X} \times \mathbb{R}$.
- Our statistical model is that the observations (x_i, y_i) are IID.
- We are given a training sequence $z_1, \dots, z_n \in \mathbf{Z}$ and a test object x_{n+1}; our goal is to predict its label y_{n+1}.

Randomized predictive systems (1)

A function $Q : \cup_{n=1}^{\infty}(\mathbf{Z}^{n+1} \times [0,1]) \to [0,1]$ is called a randomized predictive system (RPS) if:

R1a The function $Q(z_1, \dots, z_n, (x_{n+1}, y), \tau)$ is monotonically increasing in both y and τ.

R1b
$$\lim_{y \to -\infty} Q(z_1, \dots, z_n, (x_{n+1}, y), 0) = 0,$$
$$\lim_{y \to \infty} Q(z_1, \dots, z_n, (x_{n+1}, y), 1) = 1.$$

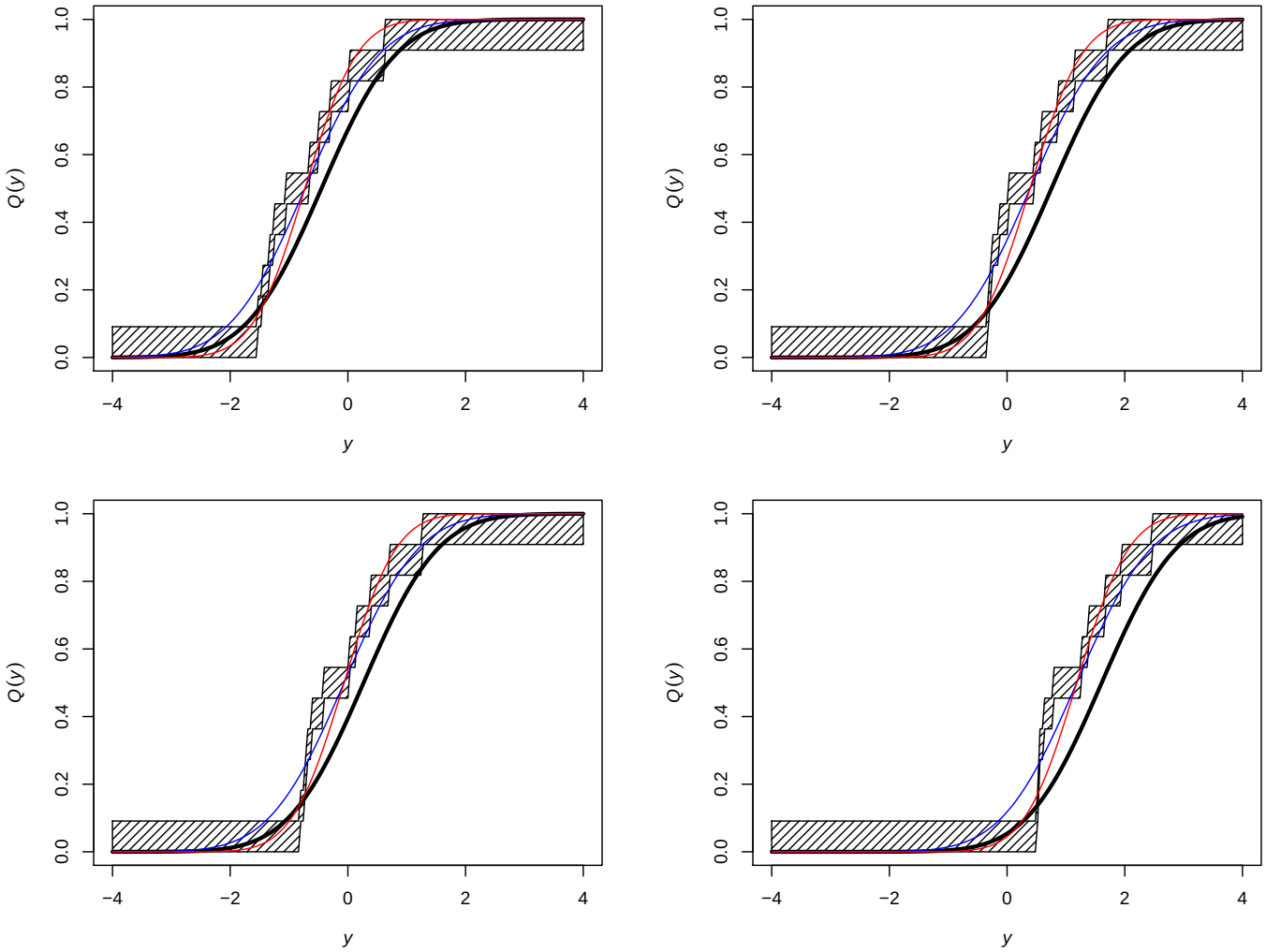
R2 The distribution of Q is uniform: $Q(z_1, \dots, z_n, z_{n+1}, \tau) \sim U$ when $z_1 \sim P, \dots, z_{n+1} \sim P$ and $\tau \sim U$ (all independent), $U = U[0,1]$ being the uniform distribution on $[0,1]$.

Randomized predictive systems (2)

- The uniformity allows us to extract prediction intervals with guaranteed coverage.
- This requirement was introduced by Shen et al. (2017) and Schweder and Hjort (2016).
- Given a training sequence $z_1, \dots, z_n$ and a test object x_{n+1}, the randomized predictive system outputs the predictive distribution (function)

$$Q_n : y \in \mathbb{R} \mapsto Q(z_1, \dots, z_n, (x_{n+1}, y), \tau).$$

An example (“LSPM”)



The example

- The shaded area: a “thick distribution function”.
- Typical vertical thickness: $1/(n+1)$ (except for n points), where n is the length of the training sequence ($n = 10$ in the picture).
- We can regard (y, τ) as the coordinate system for the shaded area.

Conformal prediction (1)

- A conformity measure is a measurable function $A : \cup_{n=1}^{\infty} \mathbf{Z}^{n+1} \to \mathbb{R}$ that is invariant with respect to permutations of the first n observations: for any n and any permutation π of $\{1, \ldots, n\}$,

$$A(z_1, \ldots, z_n, z_{n+1}) = A\left(z_{\pi(1)}, \ldots, z_{\pi(n)}, z_{n+1}\right).$$

- Intuitively, A measures how large the label y_{n+1} in z_{n+1} is.
- A simple example:

$$A(z_1, \ldots, z_{n+1}) := y_{n+1} - \hat{y}_{n+1}.$$

Conformal prediction (2)

The conformal transducer determined by a conformity measure A is

$$\begin{aligned} &Q(z_1,\dots,z_n,(x_{n+1},y),\tau) \\ &\quad := \frac{1}{n+1}\left|\left\{i=1,\dots,n+1 \mid \alpha_i^y < \alpha_{n+1}^y\right\}\right| \\ &\qquad + \frac{\tau}{n+1}\left|\left\{i=1,\dots,n+1 \mid \alpha_i^y = \alpha_{n+1}^y\right\}\right|, \end{aligned}$$

where for each $y \in \mathbb{R}$ the corresponding conformity scores are:

$$\begin{aligned} \alpha_i^y &:= A(z_1,\dots,z_{i-1},z_{i+1},\dots,z_n,(x_{n+1},y),z_i), \\ &\qquad i=1,\dots,n, \\ \alpha_{n+1}^y &:= A(z_1,\dots,z_n,(x_{n+1},y)). \end{aligned}$$

Conformal prediction (3)

- A conformal predictive system (CPS) is a function which is both a conformal transducer and a randomized predictive system.
- Any conformal transducer defines the central conformal predictor

$$\Gamma^\epsilon(z_1, \dots, z_n, x_{n+1}, \tau) := \{y \in \mathbb{R} \mid Q(z_1, \dots, z_n, (x_{n+1}, y), \tau) \in (\epsilon/2, 1 - \epsilon/2)\},$$

where $\epsilon \in (0, 1)$ is a given significance level.

Conformal prediction (4)

- The standard property of validity for conformal transducers is that the values $Q(z_1, \ldots, z_{n+1}, \tau)$ are distributed uniformly on $[0, 1]$.
- This property:
  - coincides with requirement R2 in the definition of an RPS
  - implies that the probability of error,

$$y_{n+1} \notin \Gamma^\epsilon(z_1, \ldots, z_n, x_{n+1}, \tau),$$

for the central conformal predictor is ϵ at any significance level ϵ.

LSPM (Least Squares Predictive Machine)

- The ordinary LSPM is defined to be the conformal transducer determined by

$$A(z_1, \dots, z_{n+1}) := y_{n+1} - \widehat{y}_{n+1},$$

where $\widehat{y}_{n+1}$ is the prediction for y_{n+1} computed using Least Squares from x_{n+1} and $z_1, \dots, z_{n+1}$ (including z_{n+1}) as training sequence.

- The deleted LSPM is based on the conformity measure

$$A(z_1, \dots, z_{n+1}) := y_{n+1} - \hat{y}_{n+1},$$

where $\widehat{y}_{n+1}$ is replaced by the prediction $\hat{y}_{n+1}$ for y_{n+1} computed using Least Squares from x_{n+1} and $z_1, \dots, z_n$ as training sequence.

A third version of the LSPM

- The version that is most useful for our purposes is the "studentized LSPM", which is halfway between ordinary and deleted LSPM.
- The ordinary and deleted LSPM are not RPS: they do not always satisfy R1a.
- However, we will see that this can happen only in the presence of "high-leverage points".
- And the studentized LSPM is an RPS.

More notation

- Let $\bar{h}_i$, $i = 1, \ldots, n+1$, be the diagonal elements of the hat matrix for $x_1, \ldots, x_{n+1}$.
- Huber proposed to regard points x_i with $\bar{h}_i > 0.2$ as influential.

Ordinary LSPM

Proposition

The function

$$Q_n(y,\tau) := Q(z_1,\dots,z_n,(x_{n+1},y),\tau)$$

output by the ordinary LSPM is monotonically increasing in y provided $\bar{h}_{n+1} < 0.5$.

Proposition

The above proposition ceases to be true if the constant 0.5 *in it is replaced by a larger constant.*

Deleted LSPM

Proposition

The function Q_n output by the deleted LSPM is monotonically increasing in y provided $\max_{i=1,\dots,n} \bar{h}_i < 0.5$.

Proposition

The above proposition ceases to be true if the constant 0.5 *in it is replaced by a larger constant.*

Studentized LSPM

The studentized LSPM is based on the conformity measure

$$A(z_1, \dots, z_{n+1}) := \frac{y_{n+1} - \hat{y}_{n+1}}{\sqrt{1 - \bar{h}_{n+1}}}.$$

This residual is intermediate between ordinary and deleted: a standard representation for the deleted residual is

$$y_{n+1} - \hat{y}_{n+1} = \frac{y_{n+1} - \hat{y}_{n+1}}{1 - \bar{h}_{n+1}}.$$

Proposition

The studentized LSPM is an RPS and, therefore, a CPS.

Discussion

- All three versions of the LSPM become the Dempster–Hill procedure when the xs are omitted.
- The ordinary LSPM is a very natural generalization of Dempster–Hill (details omitted).
- All three versions of the LSPM are asymptotically very close; the efficiency results (to be discussed later in the talk) are first proved for the ordinary LSPM and then extended to the other two versions.

My plan

1. Conformal predictive distributions
2. Validity and efficiency of CPSs
   - Validity in the online mode
   - Asymptotic efficiency of the LSPM
   - Universally consistent CPS
3. Conclusion and further details

Prediction in the online mode

Protocol

ONLINE MODE OF PREDICTION

Nature generates an observation $z_1 = (x_1, y_1)$
from a probability distribution P
for $n = 1, 2, \ldots$ **do**
Nature independently generates a new observation
$z_{n+1} = (x_{n+1}, y_{n+1})$ from P
Forecaster announces Q_n, the predictive distribution
for y_{n+1} based on $(z_1, \ldots, z_n)$ and x_{n+1}
set $p_n := Q_n(y_{n+1}, \tau_n)$, where $\tau_n \sim U$ independently
end for

Property of validity in the online mode

In the online mode we can strengthen condition R2 as follows:

Theorem

In the online mode of prediction (in which $(z_i, \tau_i) \sim P \times U$ *are IID), the sequence* $(p_1, p_2, \ldots)$ *is IID and* $(p_1, p_2, \ldots) \sim U^\infty$*.*

- This makes conformal predictive distributions a frequentist procedure.
- Merely correct coverage probabilities do not make a procedure frequentist!

A research programme

- Conformal predictors (and CPSs) have a property of validity under the general IID model.
- A natural question is whether, in situations where narrow parametric or even Bayesian assumptions are also satisfied, we lose a lot when relying only on the assumption of IID observations.
- This question was asked independently by Evgeny Burnaev (in September 2013) and Larry Wasserman.
- It has an analogue in nonparametric hypothesis testing: e.g., a major impetus for the wide-spread use of the Wilcoxon rank-sum test was Pitman's discovery in 1949 that even in the situation where the Gaussian assumptions of Student's t-test are satisfied the efficiency ("Pitman's efficiency") of the Wilcoxon test is still 0.95.

The parametric assumption

- The standard Gaussian linear model with additional assumptions; $\mathbf{X} := \mathbb{R}^p$.
- Given fixed objects $x_1, x_2, \ldots$, the labels $y_1, y_2, \ldots$ are generated by the rule

$$y_i = w' x_i + \xi_i,$$

  where $w \in \mathbf{X} = \mathbb{R}^p$ and $\xi_i \sim N(0, \sigma^2)$ IID.
- We assume an infinite sequence of observations

$$(x_1, y_1), (x_2, y_2), \ldots.$$

- Plus assumptions A1–A3 on the next slide.

Additional assumptions

A1 $\|x_n\| = o(n^{1/4})$.

A2 The first component of each x_n is 1.

A3 The empirical second-moment matrix has its smallest eigenvalue eventually bounded away from 0:

$$\liminf_{n\to\infty} \lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right) > 0,$$

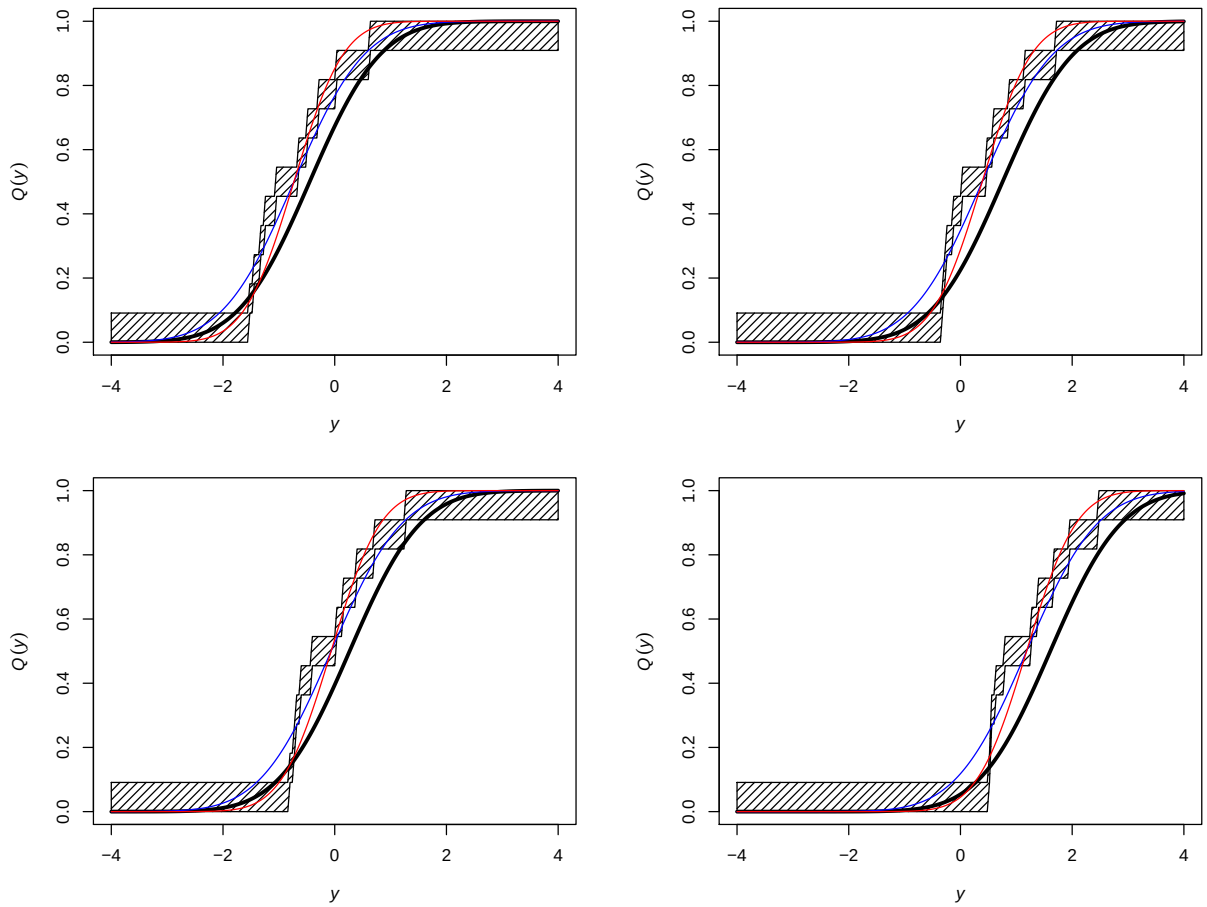
where $\lambda_{\min}$ stands for the smallest eigenvalue.

Three oracles

Intuitively, all three oracles know that the data is generated from the Gaussian linear model.

- Oracle I knows neither w nor σ.
- Oracle II does not know w but knows σ.
- Finally, Oracle III knows both w and σ.

True predictive distributions (black), conformal estimates (shaded), Oracles I (red) and II (blue)



Competing with Oracle I

Theorem

The random functions $G_n : \mathbb{R} \to \mathbb{R}$ *defined by*

$$G_n(t) := \sqrt{n}\left(Q_n(\hat{y}_{n+1} + \hat{\sigma}_n t, \tau) - Q_n^{\mathrm{I}}(\hat{y}_{n+1} + \hat{\sigma}_n t)\right)$$

weakly converge to a Gaussian process Z *with mean zero and covariance function*

$$\operatorname{cov}(Z(s), Z(t)) = \Phi(s)\left(1 - \Phi(t)\right) - \phi(s)\phi(t) - \frac{1}{2} st\phi(s)\phi(t), \quad s \le t.$$

Competing with Oracle II

Theorem

The random functions $G_n : \mathbb{R} \to \mathbb{R}$ *defined by*

$$G_n(t) := \sqrt{n}\left(Q_n(\hat{y}_{n+1} + \sigma t, \tau) - Q_n^{\mathrm{II}}(\hat{y}_{n+1} + \sigma t)\right)$$

weakly converge to a Gaussian process Z *with mean zero and covariance function*

$$\operatorname{cov}(Z(s), Z(t)) = \Phi(s)\left(1 - \Phi(t)\right) - \phi(s)\phi(t), \quad s \le t.$$

Asymptotic variance

Applying the two theorems to a fixed argument t, we obtain:

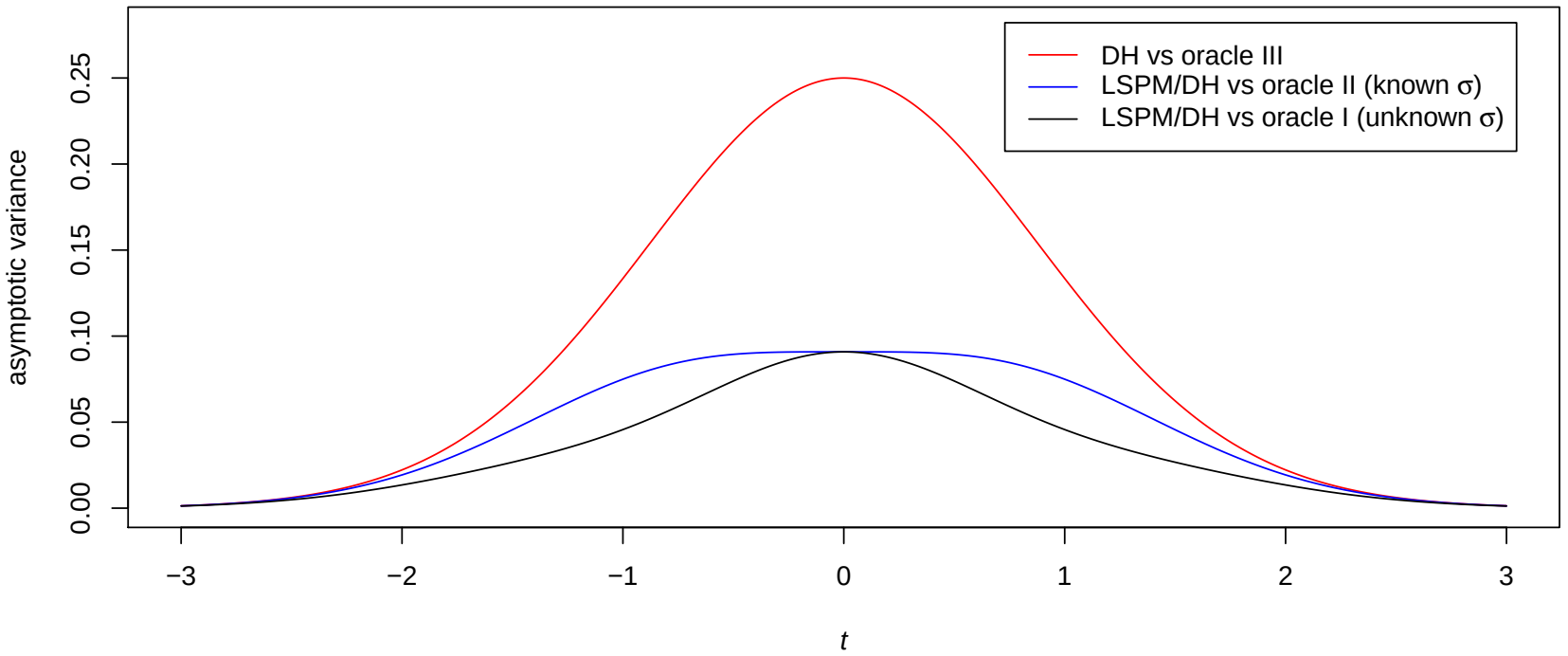
Corollary

For a fixed $t \in \mathbb{R}$,

$$\sqrt{n}\left(Q_n(\hat{y}_{n+1} + \hat{\sigma}_n t, \tau) - Q_n^{\mathrm{I}}(\hat{y}_{n+1} + \hat{\sigma}_n t)\right) \Rightarrow N\left(0, \Phi(t)(1 - \Phi(t)) - \phi(t)^2 - \frac{1}{2}t^2\phi(t)^2\right),$$

$$\sqrt{n}\left(Q_n(\hat{y}_{n+1} + \sigma t, \tau) - Q_n^{\mathrm{II}}(\hat{y}_{n+1} + \sigma t)\right) \Rightarrow N\left(0, \Phi(t)(1 - \Phi(t)) - \phi(t)^2\right).$$

Plot of asymptotic variances (DH = Dempster–Hill)



Consistency

A randomized predictive system Q is consistent for a probability measure P on $\mathbf{Z}$ if, for any bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$\int f \mathrm{d}Q_n - \mathbb{E}_P(f \mid x_{n+1}) \to 0 \qquad (n \to \infty)$$

in probability, where:

- Q_n is the predictive distribution $Q_n : y \mapsto Q(z_1, \dots, z_n, (x_{n+1}, y), \tau_n)$ (for a given τ_n)
- $\mathbb{E}_P(f \mid x_{n+1})$ is the conditional expectation of $f(y)$ given $x = x_{n+1}$ under $(x, y) \sim P$
- $z_n \sim P$ and $\tau_n \sim U$, $n = 1, 2, \dots$, are all independent

Universally consistent CPS

- The randomized predictive system Q is universally consistent if it is consistent for any probability measure P on $\mathbf{Z}$.
- This definition is based on Yuri Belyaev's definitions of related notions. (Belyaev: Moscow State University, then Umeå University.)

Theorem

Suppose the measurable space $\mathbf{X}$ is standard Borel. There exists a universally consistent conformal predictive system.

LSPM can be "kernelized" (but this does not quite achieve universal consistency).

My plan

Puzzle

- Statisticians are fond of saying that p-values are not probabilities.
- Conformal transducers output p-values.
- But now p-values coalesce into distribution functions and somehow acquire the status of probabilities.
- Do we still need Venn predictors?

Further details (1)


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


Vladimir Vovk, Jieli Shen, Valery Manokhin, and Min-ge Xie.
Nonparametric predictive distributions based on conformal prediction.
`http://alrw.net`, Working Paper 17, March 2017.
COPA 2017 Proceedings.

Further details (2)


Vladimir Vovk.
Universally consistent predictive distributions.
http://alrw.net, Working Paper 18, April 2017.


Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alex Gammerman.
Conformal predictive distributions based on kernel ridge regression.
In preparation.

Thank you for your attention!