The worst data for hierarchical log-linear models

František Matúš (Prague)

Poster at
LMS Durham Symposium
Mathematical Aspects of Graphical Models
June 30 – July 10, 2008

The hierarchical model $\mathcal{E}_{N,\mathcal{A}}$

Let N be a nonempty finite set,
$\mathcal{A}$ a family of subsets of N such that $\bigcup \mathcal{A} = N$, and
$X = \prod_{i \in N} X_i$ the Cartesian product of finite state spaces.

A probability measure (pm) Q on X is called $\mathcal{A}$-factorizable if for each $I \in \mathcal{A}$ there exists a real function ψ_I on $X_I = \prod_{i \in I} X_i$ s.t.

$$Q(x) = \prod_{I \in \mathcal{A}} \psi_I(\pi_I x)\,, \qquad x \in X\,,$$

where π_I projects x to X_I.

The set of all $\mathcal{A}$-factorizable pm's that are positive, $Q(x) > 0$ for $x \in X$, is denoted by $\mathcal{E}_{N,\mathcal{A}}$.

Information divergence from a model

The information divergence or relative entropy between pm's P, Q on X is given by

$$D(P\|Q) = \begin{cases} \sum_{x\,:\,P(x)>0} P(x) \ln \frac{P(x)}{Q(x)}\,, & \text{if } P \ll Q\,, \\ +\infty\,, & \text{otherwise,} \end{cases}$$

and the divergence of P from a model $\mathcal{E}$ by

$$D(P\|\mathcal{E}) = \inf_{Q\in\mathcal{E}} D(P\|Q)\,.$$

If P is the empirical distribution of a dataset then a miminizer Q corresponds to an MLE estimate from the data in the model $\mathcal{E}$.

The number $D(P\|\mathcal{E})$ characterizes fit of the data to the model.

The worst data

The problem of maximization

$$\max \{ D(P \| \mathcal{E}) \colon P \text{ pm on } X \}$$

goes back to Nihat Ay (2004) *Ann. Probab.*

A maximizer P admits interpretation as the empirical distribution of a bad dataset.

Example:
$\mathcal{E} = \mathrm{Bi}(n)$, $n \geqslant 3$, has the unique global maximizer $\frac{1}{2}(\delta_0 + \delta_n)$.

In general difficult,
even for 4 binary variables with all 2-way interactions.

Upper bound on the divergence from $\mathcal{E}_{N,\mathcal{A}}$

Theorem

For any pm P on X

$$D(P\|\mathcal{E}_{N,\mathcal{A}}) \leqslant \min_{I\in\mathcal{A}} \ \textstyle\sum_{i\in N\setminus I} H(\pi_i P)$$

(Shannon entropies of marginals)

Proof: induction on $|N|$, decomposition tricks, ... $\square$

As a consequence, assuming all spaces X_i of the cardinality d,

$$\max D(\cdot\|\mathcal{E}_{N,\mathcal{A}}) \leqslant \min_{I\in\mathcal{A}} \sum_{i\in N\setminus I} \ln|X_i| \leqslant [\,|N| - \max_{I\in\mathcal{A}} |I|\,]\ln d\,,$$

For 4 binary variables with all 2-way interactions
the bound $2\ln 2$ is, however, not tight.

Matroidal hierarchical models

Consider a simple connected matroid
with the ground set N of the cardinality n,
the rank function r, and the family of bases $\mathcal{A} \subseteq \binom{N}{k}$.
Let all state spaces X_i have the same cardinality d.

Theorem

If a pm P on X satisfies

$$H(\pi_I P) = r(I) \ln d \,, \qquad I \subseteq N$$

then it attains the upper bound, $D(P \| \mathcal{E}_{N,\mathcal{A}}) = [\, n - k \,] \ln d$.
The converse holds if the matroid is uniform.

This set of equalities is equivalent to saying that P is an ideal secret sharing scheme (sss) on the set of participants N with any choice of the dealer $i \in N$ and a secret of size d
(an object studied in cryptography for more than two decades).

Example: all k-way interactions, $\mathcal{A} = \binom{N}{k}$, among n variables, each taking d values. An ideal sss corresponds to an $(n-k)$-tuple of orthogonal Latin hypercubes of the size d.

CONCLUSION

Data remote to a statistical model can have
a distinct cryptographic meaning.