Human decisions in machine learning and statistical data analysis - always required, often ignored

Christian Hennig

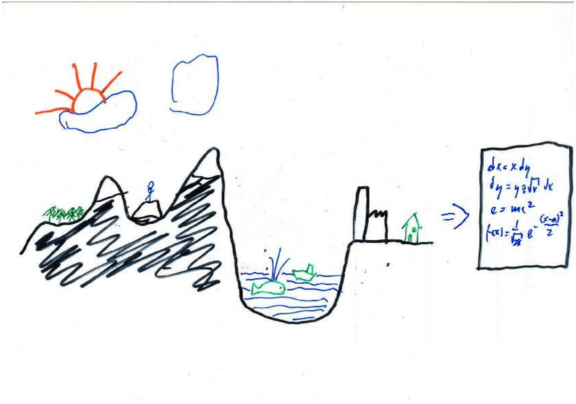
Overview

Algorithmic modelling $\leftrightarrow$ automation.
Human decisions?
What cannot be automatised in data analysis?
What are the consequences of automation?

- A constructivist view of mathematical modelling (Hennig 2010, Foundations of Science)
- Some thoughts on measurement
- On supervised and unsupervised classification (Hennig 2015, Pattern Recognition Letters)
- Beyond subjective and objective in statistics (Gelman and Hennig 2017, JRSS A)

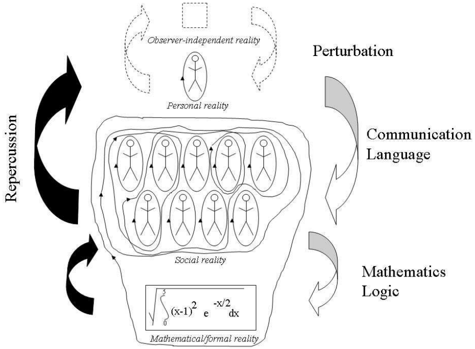
1. A constructivist view of mathematical modelling

Mathematical modelling (broad and naive):
mapping reality to mathematical objects.



Identification of items of perceived reality
with mathematical objects
and interpretation of results of mathematical operations
in terms of the items.

Constructivist view (Glasersfeld, von Foerster, Gergen,. . .)



Implications for mathematical modelling

- Science: establishing agreement in open exchange.

Implications for mathematical modelling

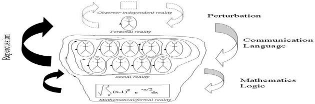
- Science: establishing agreement in open exchange.
- Mathematics is about creating a system that makes absolute agreement possible
  (but only *within mathematics*).

Implications for mathematical modelling

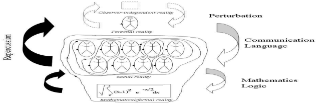
- Science: establishing agreement in open exchange.
- Mathematics is about creating a system that makes absolute agreement possible
(but only *within mathematics*).
- Mathematical modelling is not about how things are, but about how we think and communicate about them.
(Models communicate and change views of reality.)

Implications for mathematical modelling

- Science: establishing agreement in open exchange.
- Mathematics is about creating a system that makes absolute agreement possible
(but only *within mathematics*).
- Mathematical modelling is not about how things are, but about how we think and communicate about them.
(Models communicate and change views of reality.)
- It cannot be formally analysed how formal models are related to informal reality.

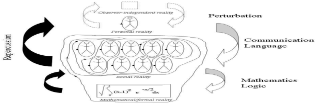


Connection to *Machine Learning*:
"machines just handle our models faster, more reliably" vs.
"machines develop models beyond our control/understanding".



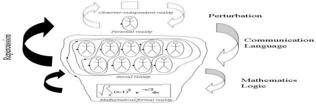
Connection to *Machine Learning*:
“machines just handle our models faster, more reliably” vs.
“machines develop models beyond our control/understanding”.

Human programmers don’t always understand implications
of their own models $\Rightarrow$ “artificial *intelligence*”.



Connection to *Machine Learning*:
"machines just handle our models faster, more reliably" vs.
"machines develop models beyond our control/understanding".

Human programmers don't always understand implications of their own models $\Rightarrow$ "artificial *intelligence*".
Ascribing intelligence to the machine
can be seen as denial of responsibility.



Connection to *Machine Learning*:
“machines just handle our models faster, more reliably” vs.
“machines develop models beyond our control/understanding”.

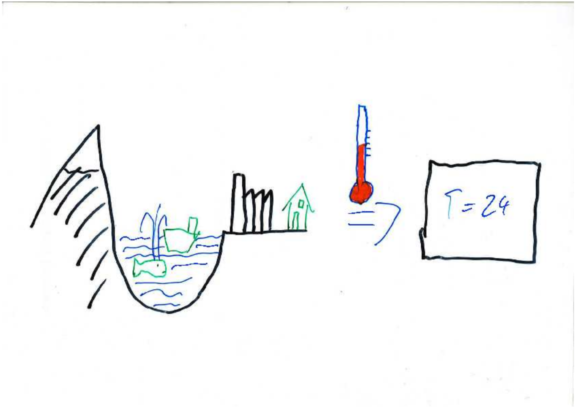
Human programmers don’t always understand implications of their own models $\Rightarrow$ “artificial *intelligence*”.
Ascribing intelligence to the machine can be seen as denial of responsibility.

Feedback on world view, intelligence concept?

2 Some thoughts on measurement
2.1 What is measurement?

The algorithms live on data, i.e., **measurements**.

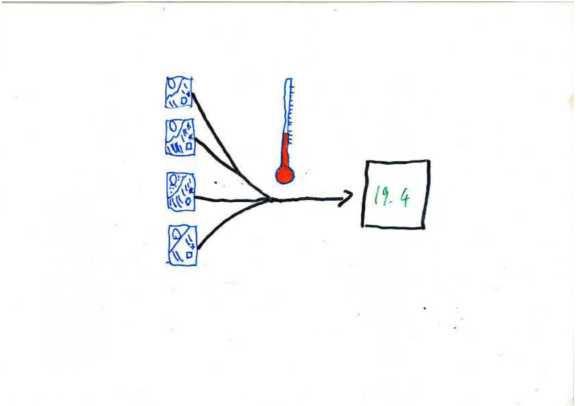
Measurement: "anchor" of modelling:
methods for assigning values to aspects of reality.



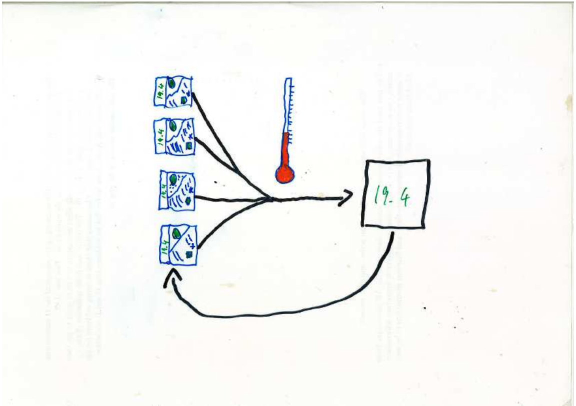
Measurement aims at descriptions of reality
that can be analysed by mathematical modelling,
and on which people can agree.

Measurement aims at descriptions of reality
that can be analysed by mathematical modelling,
and on which people can agree.

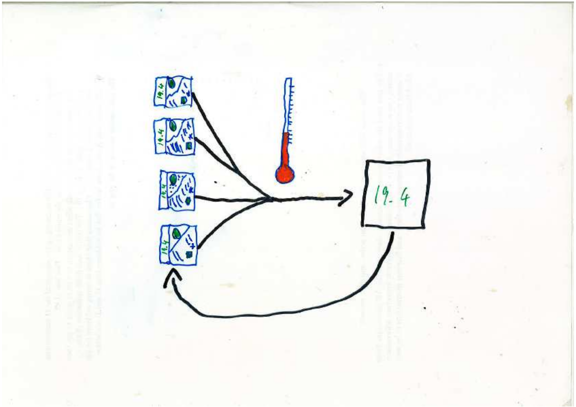
This requires a *change* of perception,
and constructive negotiation.


19.4

. . . and measurement influences perception.



. . . and measurement influences perception.



Measurement does more than just measuring.

Antony Gormley - Allotment

2.2 Measurement error

Typical statistical model for measurement T (Gauss, Laplace...):

$$T = X + \epsilon, \ \epsilon \text{ i.i.d. random (normal)}, \ E\epsilon = 0.$$

X true value, ϵ measurement error.

2.2 Measurement error

Typical statistical model for measurement T (Gauss, Laplace. . .):

$$T = X + \epsilon,\ \epsilon \text{ i.i.d. random (normal)},\ E\epsilon = 0.$$

X true value, ϵ measurement error.

X, ϵ not observable.
$E\epsilon = 0$ not verifiable.

2.2 Measurement error

Typical statistical model for measurement T
(Gauss, Laplace...):

$$T = X + \epsilon,\ \epsilon \text{ i.i.d. random (normal)},\ E\epsilon = 0.$$

X true value, ϵ measurement error.

X, ϵ not observable.
$E\epsilon = 0$ not verifiable.

Identifiability condition *defines* truth as ET!
Truth is not primary; it's derived from measurement.

Why believe in "errors"?

- Imperfect match of operation and theory/intention,
- known measurement instrument malfunction,
- known sources of "noise",
- variance of measures that "should" be the same,

Why believe in “errors”?

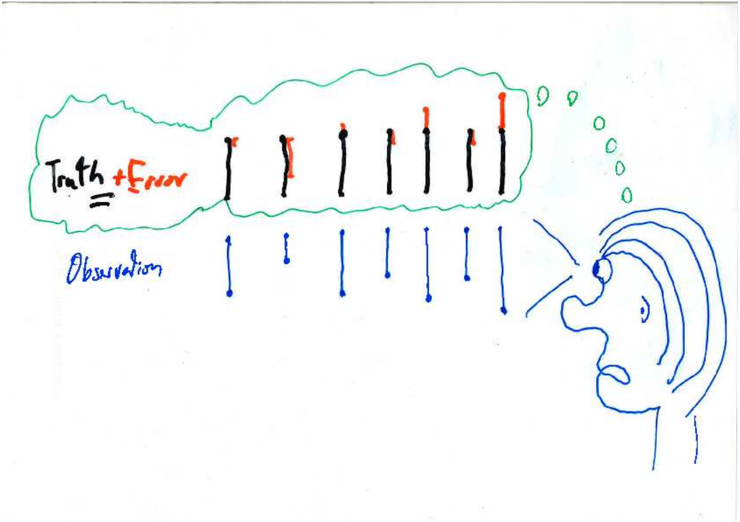
- Imperfect match of operation and theory/intention,
- known measurement instrument malfunction,
- known sources of “noise”,
- variance of measures that “should” be the same,

“Random errors” just modelling tool to deal with this.

But “error” is a loaded term.
Meaningful variation, dependence,
systematic bias may be ignored.
Model confounds “error” with “instable reality”.
Self-confirmation of “truth” concept.

But “error” is a loaded term.
Meaningful variation, dependence,
systematic bias may be ignored.
Model confounds “error” with “instable reality”.
Self-confirmation of “truth” concept.

Truth is estimated by minimising observed error.
Misinterpretation: “we know truth, error is small.”


Truth + Error
=
Observation

2.3 Validity and reliability

Validity: the procedure measures what it's supposed to measure.
Reliability: the procedure delivers a stable measurement in a stable situation.

2.3 Validity and reliability

Validity: the procedure measures what it's supposed to measure.
Reliability: the procedure delivers a stable measurement in a stable situation.

Can estimate validity and reliability from data
under assumptions of measurement error model.
But crucial assumptions cannot be checked.

Validity and reliability are unobservable because

- **validity** refers to truth
  unobservable without measurement procedure,

Validity and reliability are unobservable because

- **validity** refers to truth
  unobservable without measurement procedure,
- any two observable situations are different,
  reliability assumes them to be the same.

Both validity and reliability check instrument against researcher's constructs.

Validity: is measurement connected to external criteria in the way *the researcher thinks it should be*?

Both validity and reliability check instrument against researcher's constructs.

Validity: is measurement connected to external criteria in the way *the researcher thinks it should be*?

Reliability: is measurement stable over situations *the researcher constructs as "the same" in all relevant aspects*?

Both validity and reliability check
instrument against researcher's constructs.

Validity: is measurement connected to external criteria in the way *the researcher thinks it should be*?

Reliability: is measurement stable over situations *the researcher constructs as "the same" in all relevant aspects*?

This isn't an issue of the data alone,
researcher's world view is key.

2.4 Data in algorithmic modelling

How are the data used in algorithmic modelling connected to reality?

Often there's not much concern about meaning of data,
quality of data,
active construction of reality by
defining and processing measurements.

Data are taken for granted.

3. Supervised and unsupervised classification

Classification is a standard Machine Learning task.
Idea: observe data $\mathbf{x}_1, \ldots, \mathbf{x}_n$,
want to assign to classes $y_1, \ldots, y_n$,
e.g. $y_i \in \{1, \ldots, K\}$.

3.1 Supervised classification:

Have test data $(\mathbf{x}_1^*, y_1^*) \ldots, (\mathbf{x}_m^*, y_m^*)$
with *known* class labels.

3.1 Supervised classification:
Have test data $(\mathbf{x}_1^*, y_1^*) \ldots, (\mathbf{x}_m^*, y_m^*)$
with *known* class labels.

Some of the biggest successes of ML
are in supervised classification,
e.g. face recognition, speech recognition.

3.1 Supervised classification:
Have test data $(\mathbf{x}_1^*, y_1^*) \ldots, (\mathbf{x}_m^*, y_m^*)$
with *known* class labels.

Some of the biggest successes of ML
are in supervised classification,
e.g. face recognition, speech recognition.

Neural networks (and other methods) build extremely good
classifiers with very large amounts of training data,
if classes are well separated enough.

Supervision allows methods to be tested and compared using *cross-validation*, i.e., classify some training data ignoring the known class, see how well it's doing (misclassification rate).

Supervision allows methods to be tested and compared using *cross-validation*, i.e., classify some training data ignoring the known class, see how well it's doing (misclassification rate).

Decisions (which method? preprocessing? "tuning"?) can just be made by optimising misclassification rate.

Supervised classification is not without its problems:
May new data differ systematically from training data?
Are “known” classes reliable?
Selection bias?

Supervised classification is not without its problems:
May new data differ systematically from training data?
Are “known” classes reliable?
Selection bias?

But that’s piece of cake compared with
unsupervised classification.

3.2 Unsupervised classification (cluster analysis)

Classification with true classes unknown
and without training data.

3.2 Unsupervised classification (cluster analysis)

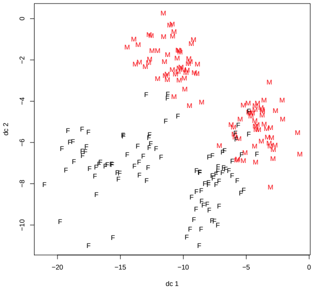
Classification with true classes unknown
and without training data.

Major difference to supervised classification:
"True" classes are implicitly constructed by the user.

Given a dataset without given classes,
there is no unique definition of true classes.

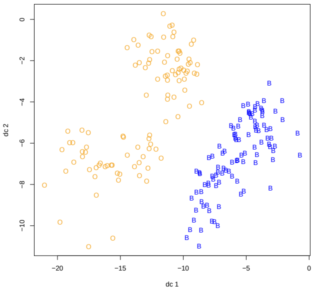
Often ignored in ML community;
for benchmarking clustering methods,
they compare misclassification rates
on data with “known” classes.

Male and female crabs.



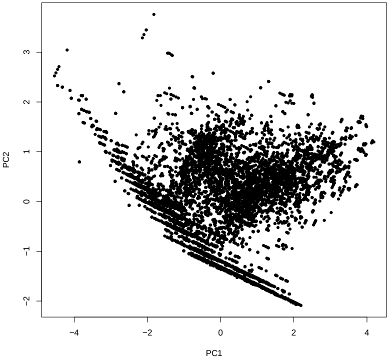
(These look quite “unclustery”.)

Different crab species.

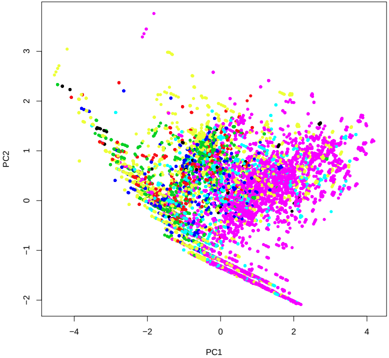


In reality there may be many valid classifications.

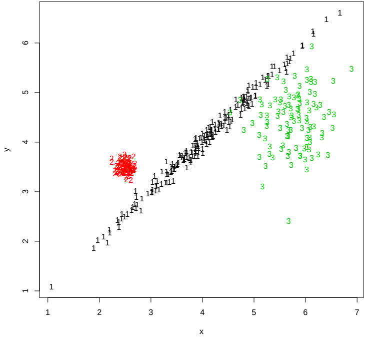
How true are the true given classes?

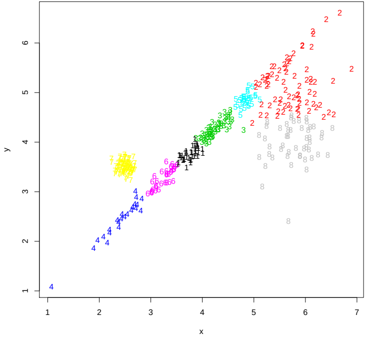


(Hennig and Liao 2013, social stratification data)



7 standard occupation classes such as “manual workers”, “managerials and professionals”, “not working”





Depending on aim of clustering
different clusterings may be preferable.

In literature, methods are advertised as
finding the “natural” clusters, assumed unique.

In literature, methods are advertised as finding the “natural” clusters, assumed unique.

Literature is not open about researcher’s need to *decide* what kind of clusters are required in given application.
Suggests data can make all the decisions.

Often clustering methods are even *advertised* saying that “no tuning/decisions are needed”.

Researcher needs to “construct” cluster concept:
by what kind of principle should clusters be separated?

(Small within-cluster distances, separation,
probability mixture components, centroids etc.)

E.g., need to specify how the idea of “social class”
is connected to income, employment, housing,. . .

Big issue: Automatic methods give results. . .

Big issue: Automatic methods give results. . .

but involved cluster concepts are prechosen,
often badly adapted to problem at hand,
will generate meaningless artifacts,
but results may be widely used and trusted.

In order to get cluster analysis
(and statistics in general) going,
researchers need to make some decisions.

In order to get cluster analysis
(and statistics in general) going,
researchers need to make some decisions.

But many tend to avoid or hide them, because

- they want to be objective,

In order to get cluster analysis
(and statistics in general) going,
researchers need to make some decisions.

But many tend to avoid or hide them, because

- they want to be objective,
- they want to appear objective,

In order to get cluster analysis
(and statistics in general) going,
researchers need to make some decisions.

But many tend to avoid or hide them, because

- they want to be objective,
- they want to appear objective,
- they don't know how to make these decisions and don't trust themselves when making them.

In order to get cluster analysis
(and statistics in general) going,
researchers need to make some decisions.

But many tend to avoid or hide them, because

- they want to be objective,
- they want to appear objective,
- they don't know how to make these decisions and don't trust themselves when making them.

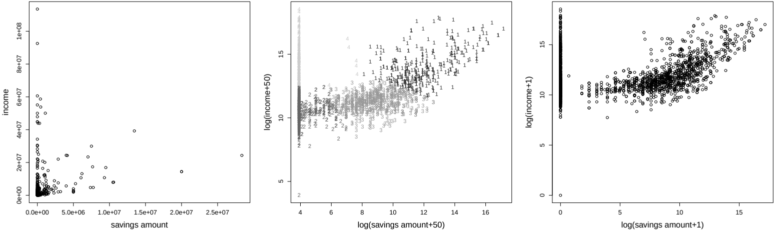
How can the algorithms make better decisions then?

4. Beyond subjective and objective in statistics

4.1 Decisions in statistics and data analysis:

- data preprocessing, transformations, representation
- variable selection, variable roles, dimension reduction
- choice of model and method, method tuning (i.e., mean vs. α-trimmed mean)
- missing values, outliers
- priors in Bayesian statistics

Impact of transformations and tuning can be huge!



4.2 Avoidance of decisions

needed for ML style automation,
but also practised by scientists.

- Prefer tuning-free methods (mean vs. α-trimmed).

4.2 Avoidance of decisions
needed for ML style automation,
but also practised by scientists.

- Prefer tuning-free methods (mean vs. α-trimmed).
  But mean is affected by outliers and
  dealing with outliers requires tuning

4.2 Avoidance of decisions

needed for ML style automation,
but also practised by scientists.

- Prefer tuning-free methods (mean vs. α-trimmed).
  But mean is affected by outliers and
  dealing with outliers requires tuning
- Use most popular/established methods
  or advertised as universal or objective (objective Bayes).

4.2 Avoidance of decisions
needed for ML style automation,
but also practised by scientists.

- Prefer tuning-free methods (mean vs. α-trimmed).
  But mean is affected by outliers and
  dealing with outliers requires tuning
- Use most popular/established methods
  or advertised as universal or objective (objective Bayes).
  Need to ignore specifics of the situation.

- Just do something and don't mention it.

- Just do something and don’t mention it.
  Oh yeah!

- Just do something and don't mention it.
  Oh yeah!
- Try to automate decisions/estimate from the data.

- Just do something and don't mention it.
  Oh yeah!
- Try to automate decisions/estimate from the data.
  Doable in supervised problems, but selection bias, and some information is not in the data, e.g., "interpretative distance" between zero and low savings, meaning of outliers etc..

4.3 Our attitude to science and objectivity

Science aims at agreement about reality; -
tries to achieve observer-independence.
In this sense objectivity is central.

4.3 Our attitude to science and objectivity

Science aims at agreement about reality; -
tries to achieve observer-independence.
In this sense objectivity is central.

But agreement in free exchange can only come from respecting individual subjective perspectives.

4.3 Our attitude to science and objectivity

Science aims at agreement about reality; -
tries to achieve observer-independence.
In this sense objectivity is central.

But agreement in free exchange can only come from respecting individual subjective perspectives.

Objectivity requires generalisation
from one situation to another;
often relies on ignoring differences.

Fundamental tension between ideal of objectivity
and subjectivity of human observers,
between attempted generality of knowledge
and specifics of every situation
is basic condition of science.

4.4 A list of virtues

Instead of branding something “objective” it seems better to more precisely discuss which virtues an approach/a study has.

Also, acknowledge more explicitly how subjectivity is indispensable; and which “subjective virtues” to achieve.

Virtues connected to objectivity

1. Transparency
    a. Clear and unambiguous definitions of concepts (*),
    b. Open planning and following agreed protocols (o),
    c. Full communication of reasoning, procedures, and potential limitations (-);

(*) Clear virtue of algorithmic modelling.
(o) Algorithmic modelling may help if done right.
(-) Algorithmic modelling may be counterproductive.

2. Consensus

a. Accounting for relevant knowledge and existing related work,
b. Following generally accepted rules where possible and reasonable (o),
c. Provision of rationales for consensus and unification;

3. Impartiality
    a. Thorough consideration of relevant and potentially competing theories and points of view,
    b. Thorough consideration and if possible removal of potential biases (o),
    c. Openness to criticism and exchange (-);

4. Correspondence to observable reality
    a. Clear connection of concepts and models to observables (o),
    b. Clear conditions for reproduction, testing, and falsification (o).

Virtues connected to subjectivity

1. Awareness of multiple perspectives (o)

Virtues connected to subjectivity

1. Awareness of multiple perspectives (o)
2. Awareness of context-dependence
    a. Recognition of dependence on specific contexts and aims (-),
    b. Honest acknowledgement of the researcher's position, goals, experiences, and subjective point of view (-).

Connecting objectivity and subjectivity

Investigation of stability

a. Consequences of alternative decisions and assumptions that could have been made in the analysis (o),
b. Variability and reproducibility of conclusions on new data (o).

5. Conclusion

- *Much* goes into statistics and data analysis that cannot be automatised.

5. Conclusion

- *Much* goes into statistics and data analysis that cannot be automatised.
- Repercussion between statistical models/measurement and modeller's/user's world view.

5. Conclusion

- *Much* goes into statistics and data analysis that cannot be automatised.
- Repercussion between statistical models/measurement and modeller's/user's world view.
- Trusting Machine Learning often comes with ignorance for required decisions, denial of responsibility.

5. Conclusion

- *Much* goes into statistics and data analysis that cannot be automatised.
- Repercussion between statistical models/measurement and modeller's/user's world view.
- Trusting Machine Learning often comes with ignorance for required decisions, denial of responsibility.
- Ignoring this we still get *results* from algorithms, but we're taking a big risk!

References:

A. Gelman and C. Hennig (2017) Beyond subjective and objective in statistics. To appear in *Journal of the Royal Statistuical Society A* (with discussion). www.rss.org.uk/Images/PDF/publications/2017/Gelman-Hennig-April-17.pdf

C. Hennig (2010) Mathematical Models and Reality - a Constructivist Perspective. *Foundations of Science* 15, 29-49.

C. Hennig (2015) What are the true clusters? *Pattern Recognition Letters* 64, 53-62.

C. Hennig and T. F. Liao (2013) How to find an appropriate clustering for mixed type variables with application to socioeconomic stratification. *Journal of the Royal Statistical Science C* 62, 309-369 (with discussion).