Journal of CHEMOMETRICS

Published online in Wiley Online Library: 5 October 2015

(wileyonlinelibrary.com) DOI: 10.1002/cem.2740

Quantitative Big Data: where chemometrics can contribute

Harald Martens*

1. INTRODUCTION

1.1. A lot to learn, a lot to give

The editors of J. Chemometrics have been kind enough to invite me to outline some directions, which I think young chemometricians could find interesting for the future. This perspective paper is my response, in the form of an essay that summarizes how I see the field today. It builds on mistakes I have made and things I have learnt in my past 43 years of data analytic R&D, in food science and agriculture, in telecommunication, analytical chemistry and biomedicine, in different universities and companies, in many countries [1–6].

This is not a review the field of Chemometrics. Instead, I present topics that I find interesting, challenging and not yet fully developed, hopefully within a coherent perspective. My biased selection of references was chosen to illustrate and expand on this perspective, linking to previous literature for further reading.

The paper is primarily written for chemists and bio-medical scientists but readers from other fields may also find something of interest, for all sciences now face the same two problems: How to convert the future's Big Data from a burden to a boon, and how to combine theory and practice.

How is the Real World joined together? What makes it tick? How can we know that, and what makes *us* tick? These are topics addressed by chemometricians and by our multivariate soft-modeling parents and siblings in fields like psychometrics, sensometrics, morphometrics etc. Since Svante Wold and Bruce Kowalski started it in the mid 70s, Chemometrics has grown to become a mature field of science, with an extensive literature and increasingly interacting with other fields.

I believe the field of Chemometrics has a lot to learn from other disciplines—mathematics, statistics or computer science. Our heads are filled with domain-specific knowledge, so many of us are lousy in mathematics except for simple linear algebra of the type $\mathbf{A} = \mathbf{B} \times \mathbf{C} + \mathbf{D}$, and pretty bad in statistics except for experimental design and simple cross-validation. Our multivariate methods—at least the ones that I know well—are powerful and versatile but they are linear and static. To describe the real world realistically, I think they need further development, e.g. with respect to temporal dynamics and feedback, spatial distribution, heterogeneity and nonlinearity. And most of our successful software packages now need fresh impulses from computer science and cognitive science.

On the other hand, Chemometrics also has a lot to give to other disciplines. Our internal culture favors warm-hearted cooperation, rather than cutthroat competition. Academically, we want Real-World science, so we cherish a humble but aggressive attitude and despise passive arrogance. Our methods and approaches allow us to handle big data tables without being overwhelmed. We do not limit ourselves to over-simplified versions of "The Scientific Method"—the testing of hypotheses and searching for p-values—we also know how to listen and learn from real-world data, and leap forward from there. We have Moore's Law on our side, but use the increasing computer power differently from many other fields: we tend to avoid the alienating "black box" modeling of machine learning and the scientific hubris of overly confident causal mechanistic modeling. Instead, we analyze big, real-world data sets with transparent data modeling methods that help us overview complex systems:

Our main data modeling tools—"factor-analytic" decomposition methods like Principal Component Analysis (PCA) and Partial Least Squares Regression (PLSR) and the many extensions thereof—help us find, quantify and display the essential relationships—expected or unexpected—within and between data tables. These transparent, open-ended methods reveal the systematic relationships, not as magic, but for the eyes of scientist to see. Because meaningful data-driven modeling requires good data, we insist on representative sampling and pragmatic, understandable statistical assessments. Thereby we can get a good grip on the complexity of the real physical world. We can also use this approach to study the behavior of humans, and even of complex mathematical models.

In the following, I present my view of Chemometrics as a science culture for the future, and outline a philosophical framework for that. I then describe some topics for future work in the field.

1.2. Quantitative Big Data

Why should eager young scientists learn data modeling tools from Chemometrics and related cultures, and try to improve them? In my opinion: (i) because science needs better data analysis, (ii) because to understand what a data set means, domain-specific background knowledge is also required, and (iii) because real-world data modeling is good fun, gives good jobs and solves real problems—all at the same time.

The fun comes with the discovery process: when analyzing a new table of data, the first, rough PCA is always exciting: what

* *Correspondence to: Martens, Harald, Norwegian U. of Science and Technology, Department of Engineering Cybernetics, Gløshaugen, Trondheim 7034, Norway. E-mail: harald.martens@ntnu.no*

H. Martens
Norwegian U. of Science and Technology, Department of Engineering Cybernetics, Gløshaugen, Trondheim 7034, Norway

is hiding behind the numbers? The good jobs come because there are not enough willing and able statisticians or other professional data analysts, so chemometric competence is sought after, at least in industry and applied R&D fields. And we have real problems, all right: in a modern world, values and wealth are more easily shared, but so are microbes and mines. Unfortunately, the human brain does not develop equally fast. We face real problems concerning environment and climate, poverty and war, health and nutrition. Some of these problems are because of conflict of interests, power misuse and self-serving cultural arrogance, resulting in serious environment problems, bloody wars and glaring injustice. These can hardly be solved by science but by ethics and politics. However, many problems arise because of lack of foresight and overview, lack of human communication and lack of sensible solution alternatives. Therefore we need better tools to understand ourselves, our cultures and the real world. For a start, we need a lot of good data.

And data we get, indeed. We are witnessing an explosion in the amounts of available data. Not just one explosion like a bomb—and not like the continuous, explosive heating in the sun but an incomprehensible, exponentially growing deluge. Some parts of the data flow are diverse, flawed and unsystematic but a lot of it represents precise, accurate, systematic high-dimensional measurements—what I call here Quantitative Big Data and these data need to be digested.

Our computer capacity keeps increasing—both in respect to storage space, memory and CPU. Unfortunately, saving ever-increasing files of raw data to every-increasing computer disk racks is not the final answer. The more data we just store, the less information we have: Unless we can also interpret these data we are overwhelmed by them—to the extent that we may be happy with disk crash. Whether we consider the big questions like global warming and multi-resistant bugs, or the daily chores in a competitive R&D world, we need to deal with Quantitative Big Data properly.

It will not be enough to have efficient "black box" algorithms, if we cannot understand the results. We have to keep people in the loop, and insist on predictive validity as well as domain-specific interpretability. So on the one hand we need a lot more cooperation with cognitive science colleagues in neurophysiology, social sciences and the humanities. That will require all parties to show each other more respectful curiosity and less fear. On the other hand we need a lot more mathematics and statistics but of the useful type that helps us be "approximately right instead of precisely wrong". Sadly, I do not think we need more mathematicians and statisticians, unless they change their ways.

2. THE ABYSS BETWEEN THE MATHEMATIZING CULTURES AND THE REST OF THE SCIENCES

2.1. The unspoken academic pecking order

In academia—like in the rest of society—there is a deep gap between math/statistics and the rest of us (Figure 1)—what will here be called "the math gap" for short. The amount of good measurements increases, but many scientists' competence and self-confidence in data analysis is low. A reason may be that there is a deep divide between the established math-statistics *cultures* and most other science *cultures.* This is reinforced by different internal recruitment policies and by the way math and statistics are being taught. A given mathematical formalism strikes some with its deep beauty, others with its narrow brutality. To some, classes on mathematical proofs and imaginary dice throwing are unbearably boring. Nowhere is the abyss deeper than in the bio-medical sciences; it may appear that "Bio is bio, and math is math, and never the twain shall meet". I believe part of the problem is mental laziness—from all parties involved. But part of it is also because of bad academic traditions.

Many theoreticians regard mathematics as "the lingua franca of science"—a language common for all scientists in all sciences at all times. But in practice, this only works for a minority today. Referring back to Figure 1, the unspoken academic pecking order cements the gap between math and the rest of sciences. By and large, I believe both sides make an effort to bridge the math gap—although asymmetrically, with the former usually as teachers and the latter usually as students. Unfortunately, much of this effort appears to be in vain. For symmetry, more theoreticians should later take applied courses.

The academic instinct, that globally generalizable results are more valuable than local ad-hoc results, is healthy. And I subscribe to the saying "Math is cheaper than physics". That is why I spent a good part of my career developing calibration mathematics in analytical chemistry [4]. But the maxim "Theory is better than Practice" is something else. Resorting to theory may be a person's fearful retreat from real-world complexity. The temptation to use math ability as an IQ test rather than a personality test is misplaced and self-serving. Most mathematicians and statisticians I know are very nice people. But I have met mathematicians as well as statisticians—in several countries—who expressed themselves very clearly in that arrogant way, primarily in situations where they thought they were "among their own". So for me, half of mathematics and statistics could relocate to the Faculty of Theology. The problem is that the other half is desperately needed in real-world R&D. And I am not sure where to draw the demarcation. So these fields are better left to define their own internal cultures. And they would probably not listen to us, in anyway. But I think we data modelers should continue to whimp and bark at them, like adolescent puppies with a twig, eagerly hoping they will come out and play.

The arrogant theoreticians may be a shrinking minority now, but they represent a fearsome tradition. Is this why we have wide-spread fear of math in our society today? We face a cultural and educational crisis now, with respect to mathematics and statistics. We need far better ways to handle today and tomorrow's complex society. For me that means a need for much more math and statistics on order to interpret and utilize the torrent of data coming available. That is definitely true in my own, bio-medical field. The same can probably said for many other fields, like most chemical fields, mechanical engineering, economics and various social sciences.

I am an optimist and think that Chemometrics and other fields that have successful experience with applied math and statistics can contribute to a drastically wider use of math and statistics in the future. Mathematical modeling and computational statistics can allow us to gain cognitive access to Quantitative Big Data, while limiting the risk of unwarranted bias and false discovery. But in much of science—as in society at large—math fear is rampant. So is statistical ignorance, especially in "softer" sciences like the bio-medical sector, where far too few individuals dare to cross the gap between the mathematical-statistical sciences and the other, more experiential sciences. Given life's complexity, both sides have to rely on simplifications. But as outlined in Figure 1, I think the mind-set on the two sides of the gap tends

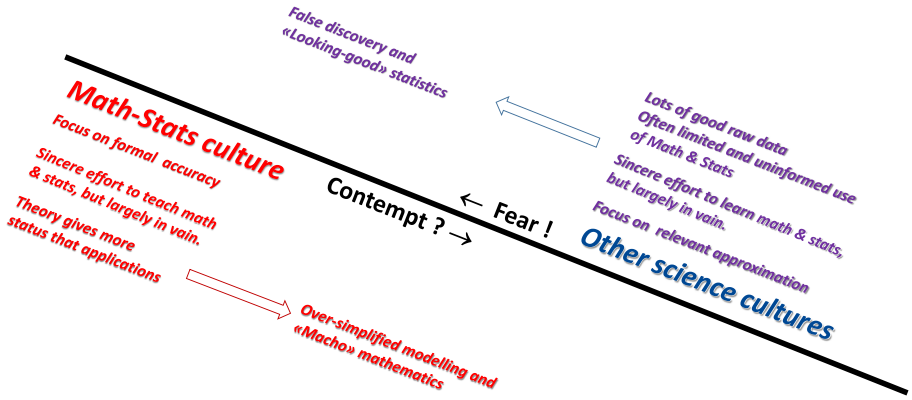


Figure 1. The math gap in science.

to be different—one side focuses on relevant approximation, the other on formal accuracy. The challenge is to bridge that gap, and I think our little field Chemometrics can contribute a lot to that.

2.2. Two roads from question to answer: induction vs deduction

Figure 2 summarizes the two main ways to come from question to answer in natural science, the data-driven way and the theory-driven way. In that sense, we chemometricians are part of an intermediate game in the philosophy of science, in that we tend to combine the two.

Chemometrics has, over the last 40 years or so developed a pragmatic science culture and a powerful set of multivariate data modeling tools, along with numerous other domain-oriented data modeling subcultures. These help the "owner" of data find, plot and interpret statistically reliable patters of co-variation in light of their background knowledge.

The common denominator of all chemometricians is to observe the Real World sincerely and with an open mind, letting the data talk to us, but at the same time trying to interpret the results in light of prior chemical knowledge and the laws of physics. Some of us have a preference for purely data-driven modeling, others for more mathematics, statistics and physics theory. All of us use background knowledge in design of experiments, in preprocessing of data and interpreting the results, and all of us look for unexpected surprises in the residuals.

Ideally, we should combine the deductive and the inductive approaches in a cyclic, type of abductive process: From a given initial hypothesis or purpose we define a rather wide observational process, measuring more properties than strictly necessary in more objects or situations that strictly necessary. To save money and to avoid locking ourselves out from surprises, we can choose broad-spectered multichannel measuring devices such as diffuse spectroscopy, and postpone the selectivity enhancement to the mathematical post-processing. To work efficiently, we use factorial statistical designs, which help us extract maximal information with minimal experimental effort.

The resulting data tables are then analyzed, without forcing the data into a straight-jacket of preconceived mechanistic models. First we use purely data drive methods such as PCA to search for gross mistakes and errors in the data, and to start getting an overview. Then we apply statistical regression methods like the PLSR to find quantitative relationships and classifications. We try various more or less physics-based preprocessing methods to improve the modeling, thereby handling variation types that we think we understand, but that can destroy the additive modeling—like separating light scattering from light absorption. And we insist on simple, but conservative statistical validation (e.g. by cross-validation/jack-knifing) and graphical interpretation at the same time, in order to avoid over-optimism.

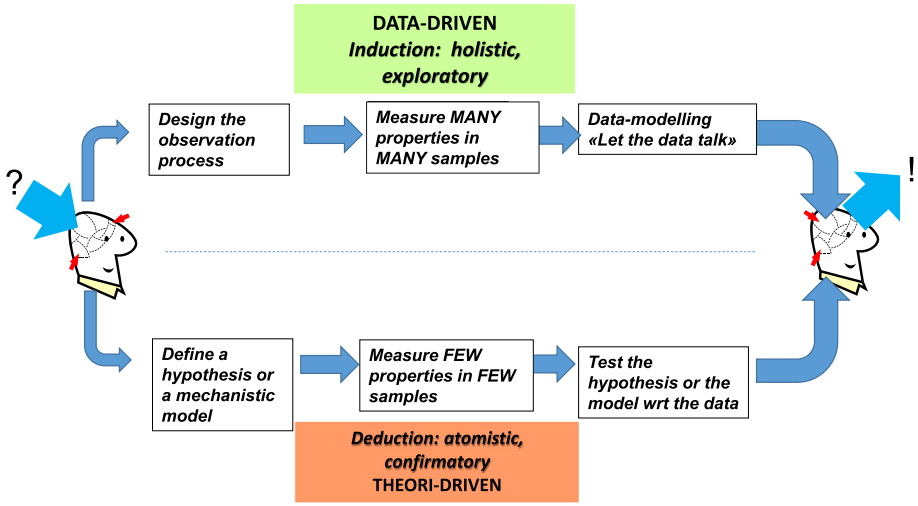


Figure 2. Two roads from question to answer.

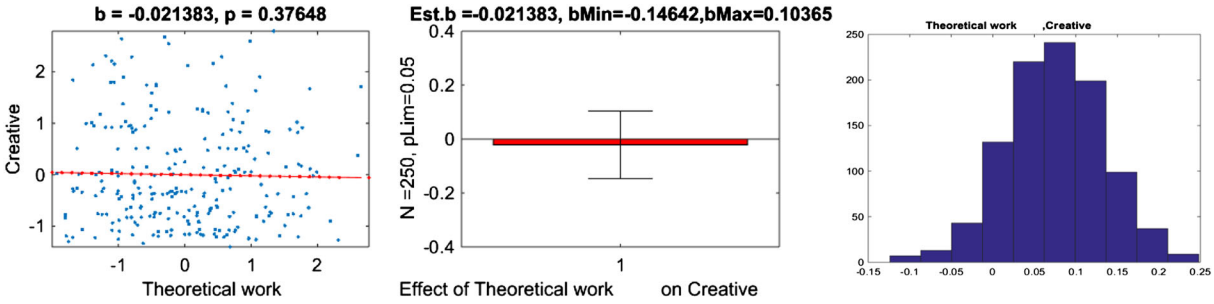


Figure 3. Are theoretically oriented people less creative than others? (a) Degree of creativity (from personality test) vs degree of theoretical work (from questionnaire) for N-250 Norwegian adults. (Both variables standardized to mean = 0 and total std.dev. = 1). (b) Estimated effect of theoretical work on creativity, with 95% confidence range. (c) Sampling uncertainty on the estimates of the effect of theoretical work on creativity: histogram of effect estimates from 1000 repeated random subsamples of N = 250 from a total of 2200 Norwegian adults [7]. *With N = 250 the data do not seem to indicate any relationship between working theoretically and being creative.*

My personal motto is: *No causal interpretation without predictive ability! No prediction without attempted causal interpretation!*

Multivariate metamodeling—the use of statistical designs and chemometric analyses to study the behavior of complex mathematical models—strengthens this bridge between induction and deduction. This will be discussed later in the paper.

Young chemometricians, faced with the future's Quantitative Big Data, will have to redefine how to combine data-driven and theory-driven modeling. It is difficult to foresee how science will develop, so they should not feel obliged to follow our old traditions.

2.3. Chemometrics for tomorrow

In the following, I shall outline some topics that I personally consider important, and hitherto unresolved. I start with the relationship between univariate statistical testing and multivariate explorative data analysis. Till now, chemometricians have focused more on statistically oriented data analysis than on mathematically oriented computer modeling. I predict that this will change. To digest and combine the enormous amounts of future scientific knowledge, we shall need the powerful tools from mechanistic modeling, too and once chemometricians gain experience with multivariate metamodeling, they will see that nonlinear dynamics and other types of mechanistic modeling is far easier than they thought. So later in the paper, I show how the borderline between data-driven "statistical" modeling and theory-driven "mathematical" modeling becomes increasingly blurred, and rightly so.

What can be done to bridge the abyss between the mathematizing science and everybody else? Is it possible to increase the number of scientists confident and capable for using mathematical modeling and statistical assessment in their work? Could it be that the pragmatic, graphically oriented data modeling is particularly suitable for particular personality types? Could it be that the Personality types most often recruited to the math and stats departments think differently from people in other disciplines, because of the theoretical nature of their subjects?

Personally, I have met a wide range of personalities in both math/stats and in more applied, empirically oriented sciences. But I believe I have observed a tendency: the math/stats people think differently from bio-people like me. Am I right? And if so, why? Were we born differently or just trained differently?

I have chosen to use this complex topic for two purposes. (i) Methodology illustration: To illustrate how traditional univariate hypothesis testing and multivariate chemometric soft modeling overlap, but differ: Significance-details vs meaningful overview. (ii) Culture critique: To contribute to the emerging academic discussion about the relationship between personality differences and science cultures. Do we chemometricians realize how different *we* are from main-stream chemists, physicists, statisticians and mathematicians?

2.4. Use and misuse of p-values

Let me pose a hypothesis: *Theoreticians are less creative than others.* If I could prove that, it would give me fat satisfaction, given my long-standing frustration about how mathematicians and statisticians teach their subjects to chemists. Let me now test my hypothesis.

Figure 3a) displays the reported degree of creativity vs degree of theoretical work for a selection of N = 250 Norwegian adults. The data were obtained from personality tests and job interviews, and is a subset of a larger study [7]. Obviously, Figure 3a) shows that people vary a lot, both in creativity and in how theoretical their work is. But is there a relationship between the two? Both the x and y variables have been standardized to a mean of 0 and a variance of 1. The regression line $\hat{y} = xb$ is shown, superimposed. Its slope $\hat{b}$ is negative; that supports my hypothesis: people who work theoretically are less creative than others. But this estimated relationship is very weak; in fact only 0.004 % of the variance in y ($r^2 = 0.00004$) can be explained linearly by x and vice versa, in this dataset. A t-test with $N - 2$ degrees of freedom[†] shows that slope $\hat{b}$ is non-significant ($p = 0.38$).

What should I conclude from that? Of course, hypotheses can never really be proven, only disproven. My hypothesis was soundly overwhelmed by its null-hypothesis: Apparently, the observed negative effect was most probably caused by "random" variation the data. What should I now conclude?

It is my impression that an uninformed use of p-values leads to needless dumbing-down in, e.g. the medical profession. The most obvious misuse is a "significant" p-value taken as a proof of an "important relationship"—or even causality [8].

Here is an important point: On the other hand, if experts just report that "the relationship is non-significant", the uninformed reader may interpret that as "there is no relationship at all". Of

[†]The regression line was obtained by fitting the mean-centered linear model $y = xb + f$ to the data [y,x] over the N = 250 subjects by ordinary least squares[(1)]. The ordinary least squares estimator is $\hat{b} = (x'x)^{-1}x'y$. A slope estimated of the line (which here corresponds to the correlation coefficient) is $\hat{b} = -0.02$. The uncertainty covariance estimate for $\hat{b}$ is $(x'x)^{-1}s_{f\hat{2}}$

course, that would be unwarranted. For one can ensure a non-significant result just by doing bad enough science—ensuring few enough and bad enough data. Why is the relationship non-significant? Is creativity truly independent of the degree of theoretical work? Or was N = 250 people too low in this case, or is the methodology too imprecise?

Equivalent situations may arise in, e.g. studies of the relationship between people's smoking and lung cancer rate, between our mobile phone use and brain cancer rate, or between young people's # of "LIKES" on social media and their suicide rate. Assume, for instance, that someone had found the risk of getting cancer from long-term extensive mobile phone use to be statistically non-significant. It is easy then to conclude: the public should rest assured that mobile (cell-) phone use is safe; case closed. That conclusion is ok if the researchers have done adequate work, so that the uncertainty of the conclusion was low, – say the effect was 1/100 000 and the upper 95% confidence limit being was 3/100 000: With 95% confidence the scientists can then claim that "No more than three in 0.0003% of the mobile users are expected to develop brain tumor because of their phone use".

But what if the researchers had checked too few, with too imprecise registrations of cancer and of mobile phone use? The estimated effect might still be 1/100 000! But the upper 95% confidence limit might now be far higher, say 3/10. So the researchers would have to report: "No more than 30% of the mobile users are expected to develop brain tumor because of their phone use". There would have been public uproar: "The uncertainty is far too high—we must finance more research!"

Thus, is not enough to report the estimated effect size and its significance stars (***, **, *, n.s.). It is important also to report the estimated uncertainty range of the observed effects, and the number of independent observations, N.

Figure 2b) shows the estimated regression coefficient $\hat{b} = -0.02$, with its estimated 95% confidence interval, which is about $< -0.15, 0.10>$. So, although no significant effect was found, it seems that the real effect might actually be as high as 0.10 and as low as -0.15.[‡]

Here is another important point: "Statistical significance" does not mean "meaningful", particularly not when N is high. The data in Figures 2a) and 2b) come from a larger study: The 250 subjects were selected at random from more than 2200 job interviews for a wide range of professions. Figure 2c) shows the distribution of the estimated connection between creativity and theoretical work, $\hat{b}$, obtained by repeating the random selection and estimation process 1000 times. It shows indeed that with only 250 respondents, the slope may be negative, but also as high as 0.25.

The corresponding correlation obtained when using all N = 2200 was found to be positive (0.073), with a 95% confidence range of $<0.07, 0.11>$. It is thus highly significant ($p < 0.001$). Still it is very small, and thus practically meaningless. Thus, in summary, Figure 2 showed that p-values are easy to misinterpret. For instance when enough data were collected, we found a slight tendency that people working theoretically are more creative than others. But it accounted only for 0.5% of the variance.

What does all the remaining unexplained variance represent? Measurement errors? Or systematic patterns of some sort?

3. SOME INTERESTING TOPICS

3.1. Personality differences: a possible explanation for the math gap?

In this project, several other variables were also recorded for each of the 2200 respondents. Can they help us understand more?

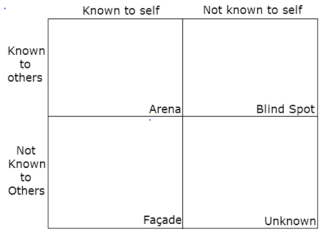
Figure 4 relates two reported personality traits, Creativity and Tidiness (Orderly, Proper), to three variables characterizing the respondents' type of work: Theoretical, Abstract and Technical work type. Then something interesting pops up: Creativity is particularly positively correlated to having Abstract work (r = 0.34), negatively related to having Technical work (r = –0.17). On the other hand, Tidiness is clearly positively correlated both to Theoretical work and to Technical work, but strongly negatively correlated to Abstract work. Apparently, there are indeed some clear structures in how people differ from each other.

The data come from an experienced psychologist's PhD thesis [7] on cognition in natural science, in particularly related to different personalities' ways of learning mathematics. There he outlined four different roads into mathematics, and how these most likely are related to differences in brain structure, causing different personality types. About 2200 Norwegian adults from a wide range of professions volunteered to take an internet-based personality test, as well as filling out a questionnaire about themselves. Before proceeding to full multivariate analysis of all the variables recorded, let me make a small detour.

The former US secretary of defense Donald Rumsfeld once made a conceptual two-by-two table of the objective reality vs what we are aware of (left side):

Known knowns	Known unknowns
Unknown knowns	Unknown unknowns

D. Rumsfeld February 12, 2002, in a news briefing on the motivation for the second US invasion of Iraq.
Source: https://en.wikipedia.org/wiki/There_are_known_knowns



The JOHARI window
Source: https://en.wikipedia.org/wiki/Johari_window

Feeling	Intuition
Sensing	Thinking

Four personality types
Source: Brovold 2014

[‡]I believe that today's misuse of p-values in applied statistics is usually not caused by professional statisticians, but by statistical autodidacts (like me), now that software for classical statistical analyses is available to everyone. Part of this, in turn, is that we amateurs do not fully understand the concept of "degrees of freedom" in ANOVA etc.

Rumsfeld is not my favorite, politically. But this table is good, for it teaches humility. And it demonstrates the benefit of looking at more than one dimension at a time. I see it as a summary of a two-component mental model, reminding us of the subjective limitations to our rationality. That applies for politicians, but for

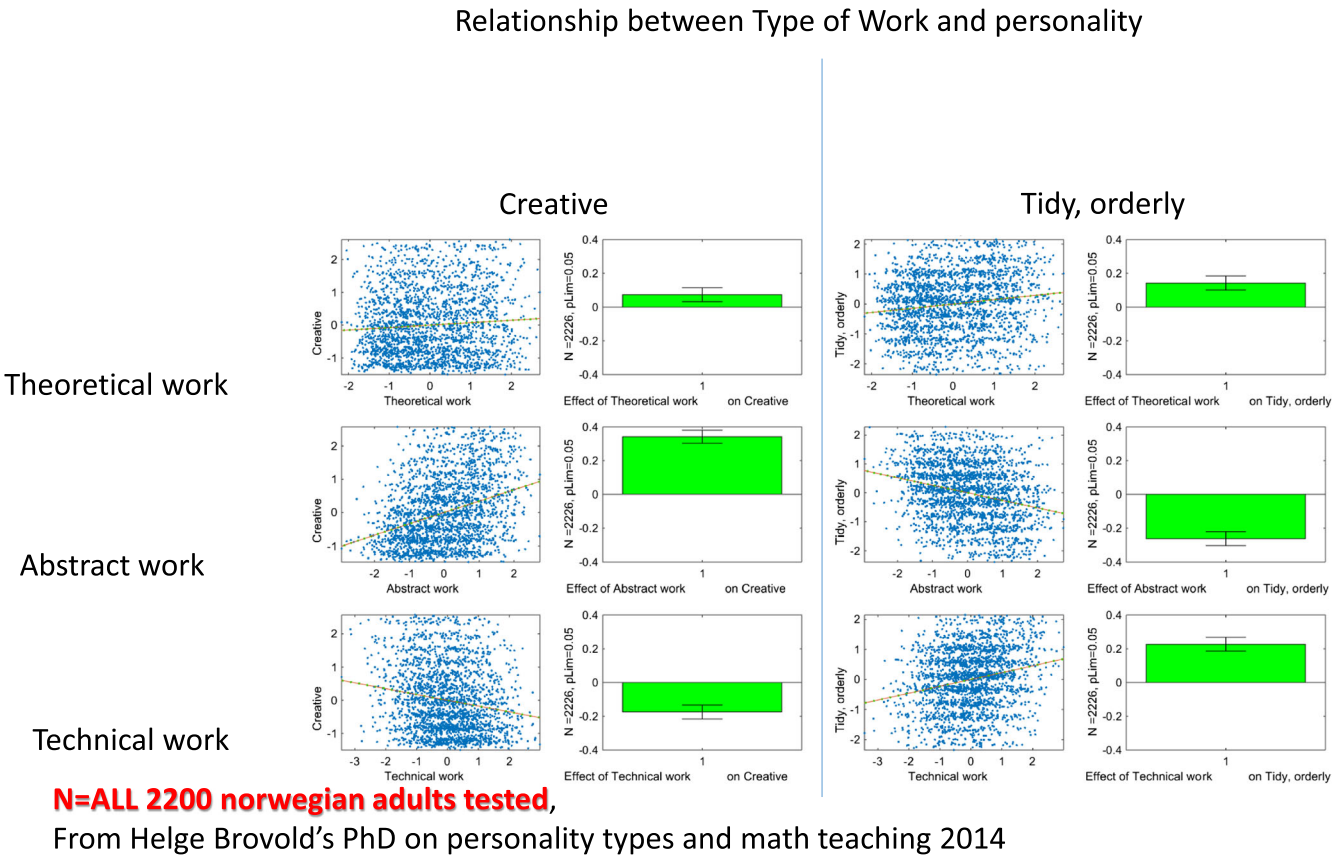


Figure 4. Relationship between reported Type of Work and measured Personality for N = 2200 Norwegian adults. Upper: degree of theoretical work. Middle: degree of abstract work. Lower: degree of technical work. Left: effect on degree of creativity. Right: Effect on the degree of being tidy and orderly. *With N = 2200 the data* [7] *indicate a statistically significant but meaninglessly small positive relationship between working theoretically and being creative.*

scientists too. Similarly, the famous "Johari window" (middle) is a two-way table or figure spanning the Self/Others axes. It reminds us that it is difficult to know oneself, and even more difficult to know each other. To cross barriers in science therefore takes an effort. In the same vein, a two-way table of dominating personality traits is also shown (right side). It represents a summary of some important personality differences that may help explain the results in Figure 4. The 2 × 2 table will now be expanded upon:

Figure 5 shows the main patterns found when analyzing the personality tests and self-assessments for the 2200 Norwegian adults. A PLSR model $Y = f(X)$ was developed for relating personality test variables *Y (in italics)* from self-reported personality traits and vocation X. Cross-validation showed two valid covariation pattern types (components), and the figure shows how they correlate to each of the X- and Y-variables. In summary, the correlation loadings show two more or less independent contrasts in the personality tests: THINKING/INTROVERT vs FEELING/EXTROVERT and INTUITIVE/CONTEXTUAL vs SENSING/DIGITAL (i.e. discretizing). The respondents' self-assessments and reported work types correlate with these two components in interesting ways: The THINKING tend to work theoretically, and to be rational, controlled and critical, while the FEELING tend to be light-hearted and work with people. The INTUITIVE tend to do abstract work, being courageous and entrepreneurial, while the SENSING tend to be tidy and humble, and do administrative or technical work.

Speculating, I would expect to find many tidy, practically oriented engineers in the lower left corner, opposite to more entrepreneurial architects, artists, media-and marketing students. Social scientists and humanities students I would expect to see in the upper left people-oriented corner, opposite to the mathematicians, statisticians, physicists and physical chemists in the lower right, theoretical.

Among data analysts, I would expect to find many full-time chemometricians in the contextual-intuitive entrepreneur corner, along with inventors of brand new modeling approaches and applications. Perhaps there is a preference, among the Intuitive and Contextual personalities (upper right) for continuous, contextual graphical mapping methods like PCA and PLSR (more about these later), while discrete cluster analysis along with various "black box" machine learning methods, may be more preferred by the SENSING/DIGITAL (i.e. discretizing) personalities (lower left)? Within statistics, I would expect traditionalists in the lower right corner, while computational statisticians could be in the lower left or the upper right corner—but probably not in upper left corner.

Figure 5 indicates what I believe is one of the main reasons for the problems in traditional math/statistics teaching to students in other fields: Different science cultures tend to attract different personality types. Mathematical proficiency has traditionally been regarded as a measure of general intelligence, to be admired. But I believe math intelligence is rather specialized. And I have met statisticians and mathematicians who arrogantly claimed to be proud about not doing applied work, only theoretical. But this is an outdated view of mathematical sciences, or at least something that I disagree with strongly. However, I have come to realize that math and science teachers are neither sadists nor snobs, usually. In fact, they often make an extraordinary

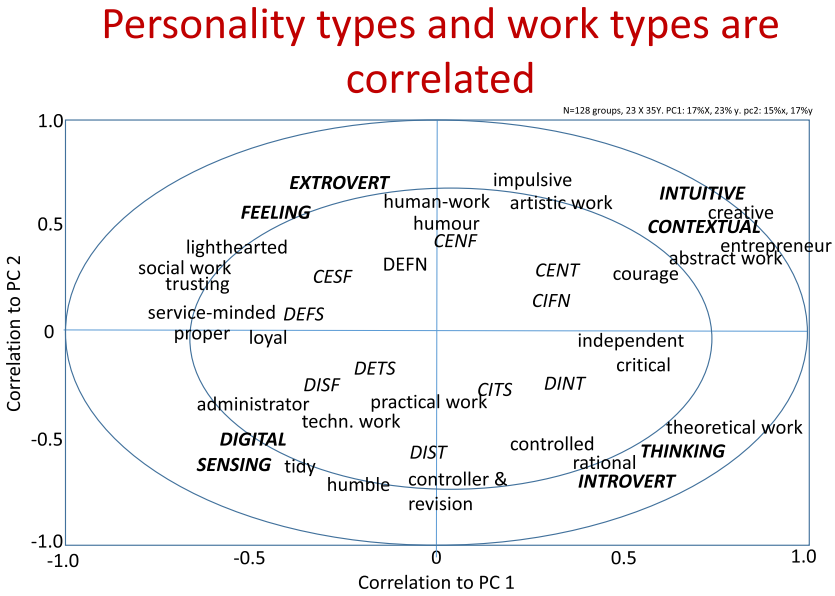


Figure 5. Different personality types ≈ different types of work; PLSR correlation loadings. Personality types (Y, *italic UPPER-CASE*) related to work types/self-assessments (X,) of 2200 Norwegian professionals [7] by sampling-balanced PLS regression. Two PLS components (PCs) were found to have predictive validity in cross-validation. The ellipses show the locus of 100% and 50% fitted variance.

effort to reach out to non-math students with their topics, and they are as frustrated as anybody about people's fear of math and loathing of statistics. Why, then, are some people fascinated by mathematics, while others hate it and fear it?

We natural scientists and technologists may not like to think about the way we think. We expect the world to behave systematically, our measurements to be precise and accurate, and our analyses to be objective, so our conclusions should be clear. We are not trained in hermeneutic concepts—how our actions yesterday affect our observations today. But fundamentally, we are in the same boat as the social scientists and even the humanities people: We are sentenced to a life as applied philosophers and dominated by our psychological preferences and limitations. But we do behave and think differently, because we are different—often in systematic ways.

It is well established that cognitive science is relevant for mathematical thinking and use. Lakoff and co-workers [9,10] even showed how important embodied metaphors are for mathematical language and concepts. Moreover, Kahneman [11] pointed out how human "thinking" has two levels of processing—a fast, intuitive, rather effortless and visually oriented capacity drawing on prior experiences and tacit knowledge, and a slow, logical, rather strenuous and often verbally oriented capacity based on explicit, formal analysis.

Personally, I have found both ways of thinking valuable in multivariate "soft" data modeling. On one hand, the intuitive envisioning is helpful for grasping one's new modeling ideas, holistically and while they are still fresh. Graphical visualization later allows me and my colleagues to overview data in light of our background knowledge. On the other hand, explicit mathematical formulation of well established theory is useful for loss-less preprocessing of chemical data. Detailed linear algebra analysis of my algorithms can reveal flaws in my thinking. For me, testing software with data simulated to contain well-defined structures is superb for finding bugs in my software implementations.

I would like to claim that theoretical distinctly oriented students are well served by traditional math and statistics teaching. But if math and statistics is primarily taught by introvert, intellectually "tidy" theoreticians, their didactic style cannot be expected to reach other main personality types. For instance, as Helge Brovold has pointed out for me, the feeling-oriented, extrovert student types (FEELING/EXTROVERT), whom we need for developing a humane science, will tend to find the theoreticians boring and the math topic meaningless. And the free-wheeling, creative entrepreneurial students are bored stiff by mathematical proofs and pre-cooked formalism—something that I can confirm from my own wasted effort as math-averse, contextually oriented biochemistry student. Today I have been professor in mathematizing fields of science in five different universities, but I have not yet been able to come through a mathematical proof awake.

Chemometrics today has a far more solid theoretical grounding that we had in the 70s and 80s. Our methods have also gained recognition among statisticians. Personally, I work mostly with mathematicians, physicists and cyberneticists these days—scientist far better in mathematical theory than me. But with its focus on relevance and real-world discovery, Chemometrics offers rich opportunities also for the intuitive, contextually oriented entrepreneurs as well as for the extrovert, feeling-oriented personality types. Our tools and our culture give us freedom to work in different fields.

3.2. Ontology: how the world is

Where should young chemometricians focus? That depends on how we think about the world. Chemometrics usually reflects a pragmatic philosophy of science [12]. Philosophizing can be useful, but is also an early sign of senility. At my age, I am entitled to it. So here is my view of what we do and what more we might do.

First, I think it is meaningful to consider that our existence has two domains—the material domain and the immaterial domain. The material domain is physical, and governed by physical laws, which are largely known even though their complex combinations bewilder us, and a lot remains unknown at the quantum level and at cosmic level. The material domain is addressed by many natural sciences, including Chemometrics. The immaterial domain is purely informational (religion, culture, concepts, language, including mathematics, and messages including telecommunication). It is addressed by information theory, and, e.g. by Sensometrics and Psychometrics.

The distinction between the material and immaterial domains is not crisp. Money, for instance, seems material one day, and the next: just a memory. But the unfortunate lack of love between natural sciences and humanities shows the distinction to be meaningful. In the present context, the genomic information content of DNA represents an example of the immaterial: This message can be held in a strand of base pairs, as a series of light signals in a DNA sequencer, as a combination of letters ATCG printed on paper or represented electronically in a PC; this can be transmitted as radio signal or over the internet, then represented as a sequence of chemical reagent concentrations in a DNA synthesizer before ending up in a DNA strand again with the same information content.

The laws governing the immaterial domain are largely unknown, I think, except for cases where exact mathematical "laws" rule, for instance how the sign and real/complex nature of the eigenvalues of the Jacobian determine the behavior of linear dynamic systems. Future chemometricians should not be afraid to combine data from the material and the immaterial domain. For even the most reductionist materialist cannot step out of the immaterial, hermeneutic spiral of purposes, and expectations.

Secondly, the physical domain itself is interesting to scrutinize more closely. The measurement revolution that we now experience in science reflects the nature of reality itself, – or at least is intended to do so. Figure 6a summarizes one very exciting aspect of Reality: It comes with only three fundamental "ontological categories": Properties, Space and Time. And in each of these categories, there may be variation along the "ordinate" (the intensity) and along the "abscissa" (the address). The former is in Figure 6a illustrated by changes in a peak's size, and the latter by changes in its position. So in total, there are $3 \times 2 = 6$ principally different modes in which Reality varies and thus can be observed[§]. The material ontology outlined in Figure 6a is the theater in which I think young chemometricians go to watch and learn. Then which show will be most fun?

3.3. Epistemology: how we observe the world

There are probably infinitely many ways to observe each of the three ontological categories in Figure 6a. Therefore, scientists have to choose—more or less subjectively—what to measure or not (Figure 6b). Based on our prior expectations, only in a tiny chosen sub-set of samples in time and space, and only a tiny sub-set of all the properties may be measured. This subjective selection, based on our more or less conscious and rational cost/benefit assessments, is a necessity and usually also an advantage. But if affordable and unique opportunities are lost because of erroneous expectations, incompetence or hidden motives, it generates and perpetuates scientific ignorance and even seem to support an erroneous bias.

Therefore, we chemometricians—like colleagues in many other fields, increasingly stress the importance of *conscious choice of "samples"*, statistically well designed and based on the existing knowledge and intuition. This planning may be done quite pragmatically [5]. But within the available cost limitations, it should at least address both *sampling* to ensure sufficient *representativity* [13], *multi-level factorial designs* to reveal possible *nonlinearities and interaction* effects, and some sort of *power assessment* to ensure that the experiment has *reasonable chance* of showing interesting results.

Moreover, the choice of measuring device(s) may benefit from the experience in, e.g. multichannel diffuse Near Infrared (NIR) spectroscopy: Measure *many* properties—preferably more than necessary, (it usually does not cost much, extra). How we then utilize these measurements depends on how we model them mathematically (e.g. linear or nonlinear structure models) and how we estimate the model's parameters statistically (e.g. based on full-rank Least Squares or Maximum Likelihood, or reduced rank approximation). Contrary to the traditional full-rank regression and discriminant analysis methods (which are still the focus of many statistics courses—alas—because of their unrealistically pretty theory), modern reduced-rank methods like Principal Component Analysis (PCA), Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) treat intercorrelations between measured variables as a stabilizing advantage, not as a "collinearity problem". Thus, by multivariate calibration, the many measured variables can often correct for selectivity problems, and multivariate outlier analysis may even reveal unexpected surprises [4]. And contrary to linear reduced-rank methods from computational statistics, like ridge regression, LASSO and elastic nets, the bilinear methods most often used in Chemometrics, cross-validated PLSR provides additional insight in terms of so-called loading—and score-plots; thereby the user's domain-specific tacit knowledge is brought into the scientific data analysis.

However, even the rank-reduced methods like PLSR suffer if too many irrelevant or noisy variables are included in the data modeling. In earlier uses of PLSR, this was usually solved by variable selection of some sort, e.g. by successive down-weighting of variables, or by jack-knifing (as part of the cross-validation, [14]). More sensitive and less "greedy" methods, like sparse PLSR, have since been developed [4,15–19].

Still, the data that we obtain will ultimately be limited by the scope of our prior expectations. And when analyzing these data, our interpretation will likewise be colored by our expectations, right or wrong. To counteract this subjectivity, exploratory analysis requires good validation methods, and this has several aspects [20]. As in science in general, data analytic validity is the extent to which a concept, measurement or conclusion is well-founded and corresponds accurately to the real world. As I see it, analysis of Quantitative Big Data requires three main validation challenges to be overcome:

(1) Cognitive validation: A data set never contains enough information in itself. Humanity's prior knowledge should be utilized, but in a soft way. If possible, one should extract the essence of the data in a way suitable for visualization and graphical validation in light of the user's more or less tacit background knowledge.
(2) Statistical validation: A data set always contains noise, and usually some mistakes as well. Traditional significance testing is intended to guard against being fooled by random errors, if used sensibly and not just to squeeze out good-looking p-values for publication. In Chemometrics' open-ended data modeling, overfitting (over-parameterization) increases the risk of false discovery. Wishful thinking and vested interests may aggravate the problem. In the exploratory data-modeling overfitting can be reduced by rank optimization [21] based on pragmatic cross-validation, or

[§]Figure 6a is inspired by the British philosopher E.J. Lowe [40], but simplified and modified to the world of Newtonian physics, where we can distinguish between time, space and properties. Lowe points out one additional ontological category, namely the presence of "gestalt"-like combinations the other ontological categories. This greatly resembles the thought model for the macroscopic level in analytical chemistry: Every chemical compound in a system has property spectrum and a concentration varying in time and space. The property spectrum is characteristic and often constant, or it varies systematically. The concentration, on the other hand, can vary more or less at random.

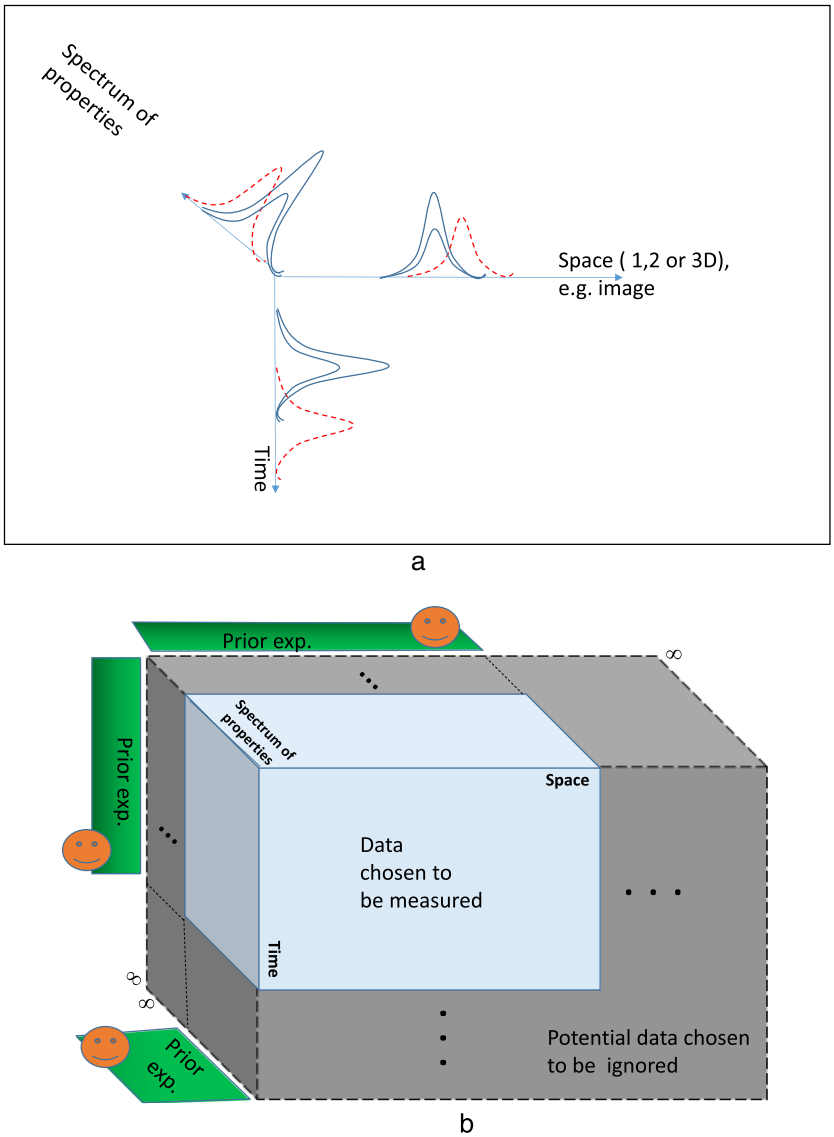


Figure 6. Thinking about the real world. (a) Ontology: What constitutes the real world? The physical reality that surrounds us is constituted by three ontological categories: Time, Space (1, 2 or 3D) and Property types (many-dimensional, e.g. light at different wavelengths). In each of these categories, there are two aspects: Position (usually thought of as an "abscissa, x-axis") and Intensity ("ordinate, y-axis"). In total: 3×2 ontological domains. An additional ontological category is represented by the "gestalts", "components" or "factors" that link the six ontological domains. (b) Epistemology: How can we observe the real world? Many modern measurement types (e.g. hyperspectral video) reflect all three ontological categories. "Properties" is here represented by "spectrum". In principle, an infinite number of time points, spatial locations and properties may be measured. In practice, funding limitations and our ignorance force us to choose only a subset of data to be measured. The rest remains in the dark. Even so, hyperspectral video etc. generates huge data arrays that require some sort of modeling.

preferably double cross-validation (cross-model validation, [22]), *in lieu* of the ideal—a sufficiently large and representative independent test set.

(3) Progressive validation: A data set never tells the whole story. The associated prior expectations—as far as we know them—should ideally be corrected for. Still, non-significant effects and outliers may later turn out to be valuable and significant results may turn out to be wrong. Therefore the scientific process of drilling into the solid Mountain of the Unknown must go on and on. Statistically valid claims should be reproduced independently. Intuitive hunches should be pursued. Solid theories are man-made and should be critiqued.

In traditional chemometrics, the three ontological modes *time, space and properties* are usually accessed in a two-way fashion: A multi-channel instrument records a set of *variables'* (a spectrum of properties) in a set of *samples* various points in time and space). But some modern instruments can observe both in time, space and property domains in a coherent fashion. For instance, a video camera allows us to record a small "spectrum" (RGB) at many pixels in space and many frames in time. Modern hyperspectral cameras extend that beyond the human visible resolution, while, e.g. medical imaging by MRI converts this from 2D to 3D. In each case, the property spectrum exists in space and time, and changes with spatial motion and temporal dynamics.

Of course the space mode itself is 3D—i.e. it has its own three internal ways. The spectrum of properties itself can also sometimes have higher internal dimensionality (fluorescence excitation/emission, chromatographic LC/MS etc.). The time mode only has one internal way. But once observed, both the

space, time and property categories may be given additional internal dimensions, e.g. by creating first and second derivatives in time, in space and along, e.g. wavelengths, both with respect to "horizontal" shifts and to "vertical" intensity changes.

In each of these many ways, the scientists (or the instrument makers) have to limit their ambitions, based on their prior expectations. Still, the amount of data from today's instruments is already staggering, and will increase drastically in the future. The culture and methodology of Chemometrics have something to contribute to the handling of such Quantitative Big Data—while we have a lot to learn, too.

Likewise, extensive computer simulations can also generate huge amounts of data. For instance, in multivariate metamodeling (see below), simulations according to large statistical designs can reveal the behavioral repertoire of complex mechanistic mathematical models. Even for Quantitative Big Data from simulations, Chemometrics has something to offer—and to learn.

Figure 7 shows a typically Chemometric approach to analyzing a data set obtained as property spectra in time and space (Figure 6b): Subspace decomposition. The data cube "**X**" is approximated by some sort of data-driven modeling, splitting it into a low number of components (sometimes called factors or estimated latent variables). Each such "gestalt" component is defined by a parameter vector (often some sort of eigenvector) in each of the data ways, representing a (linear) combination of the data in that data-way. Thus a cacophony of input data is summarized by its main underlying rhythms and harmonies, plus unsystematic errors.

In Figure 7, the conventional *bilinear* PCA model has been expanded to an N-linear model. These general N-way models (Parafac, Tucker-analysis) allow extremely compact representations of huge data files in terms of a low number of underlying patters, bundled together in N-linear "factors" or "gestalts" with manifestations in spectral properties, time and space, connected via a compact central kernel tensor. These, in turn, are suitable for practical use and graphical interpretation. The number of such components is determined empirically, e.g. by cross-validation. Expanding Svante Wold's early analysis of the bilinear PCA as a series expansion (pers.com.), I think of the N-linear decompositions as multivariate series expansions of the unknown structure in **X**.

I first met method these N-way methods in psychometrics in the 70ies, where I started, and became fascinated [1]. But after having played with them on non-ideal chemical data, I became afraid that chemists might misuse them, because their ability to reveal underlying structures unambiguously applies only for ideal data. However, Rasmus Bro and others have since convinced me of their potential for effective extraction of information, e.g. from Quantitative Big Data.

Different methods make different assumptions about the compact central kernel tensor: Parafac [23] is most strict—it assumes the kernel to be hyperdiagonal, while the Tucker 1, Tucker 2 and Tucker 3 models relax this successively, whereby the number of components may be different in different ways.

Just like PCA may be extended into PLSR, the N-linear model may be extended to match the model parameter vectors against external properties in each of these domains, as illustrated by, $\mathbf{Y}_{Spectrum}$, $\mathbf{Y}_{Time}$ and/or $\mathbf{Y}_{Space}$ in Figure 7.

A number of interesting multi-matrix and non-linear extensions of these bi-linear and N-linear decomposition methods have been developed: I think the multi-block methods of various kinds [24–27] are particularly interesting as cognitive tools to interpret the subspace structure of cross-disciplinary data. I expect the extended combination of nominal-level dynamic PLSR [28] and sparse PLSR to be particularly powerful for revealing nonlinear dynamics.

Like in all other sciences, there is a deeply subjective aspect to Chemometrics. As shown in Figure 7, the input data are first scaled by vectors of *prior scaling weights*, intended to balance the signal/noise in spectrum, time and space in various least-squares based steps in the data modeling. Samples deemed to be outliers may thus be removed. Conventionally, variables are left unscaled or standardized to equal variance. But other scaling weights may also be chosen: Variables considered to be particularly important may be forced into a data driven model by very high scaling weights. Conversely, questionable variables may be down-weighted. These scaling weights may equivalently be employed inside the iterative modeling algorithms, or explicitly as pre-processing. The latter is often preferred in Chemometrics, to save time.

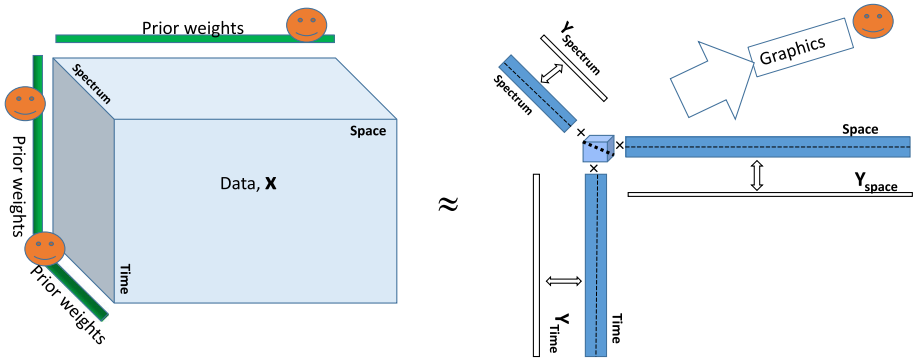


Figure 7. Data driven modeling: finding the underlying patterns in data. An N-way array **X** of high-dimensional data may often be approximated by a much simpler N-way data model. The figure illustrates three-way Tucker or Parafac data modeling wrt time, space and spectrum. Like in conventional two-way PCA, prior scaling weights, subjectively chosen by domain experts, are used to balance the signal/noise and relevance of different time points, locations and properties. The ensuing N-linear decomposition results in a few "gestalts" with properties both in time, space and spectrum. These are here collected in low-dimensional loading matrices in time, space and property types, connected by a small N-way kernel in a low-dimensional tensor model. The loadings may be related to external variables, $\mathbf{Y}_{Time}$, $\mathbf{Y}_{Space}$ or $\mathbf{Y}_{Spectrum}$, in analogy to two-way PLSR, and are suitable for graphical interpretation by the domain experts.

By these scaling weights we consciously impose our subjective "prejudices" (prior expectation of relevance and precision) on the data modeling. For later graphical analysis of the obtained loading vectors, the scaled variables may be brought back to their original units by division by their scaling weights, or plotted in terms of so-called correlation loadings. Properties totally ignored have implicit scaling weights of zero, and their modeling results of course cannot be de-scaled again. The same goes for locations, time-points or samples ignored.

I think this inversion of our subjective prior scaling weights gives a naive illustration of how, and to what limit, we can eliminate our own "pre-judices" in general [29], when trying to interpret observations in the Real World. In our unavoidable hermeneutic circle (or rather, spiral), our prior scaling may be extended from simple weight vectors to matrices of expected patterns, desired or undesired. These may be implemented in different ways, e.g. by Generalized Least Squares preprocessing (matrix multiplication by the square root of expected covariances [30]), or by statistically more advanced Bayesian analyses. Unfortunately, Evolution did not allow our brain to do matrix multiplication—otherwise we humans might have been able to use this to correct for our own cultural biases. On the other hand, to be stabile, cultures may need some shared prejudices. So perhaps we should limit our prejudice correction to computational science. But there we may have to use it, as a tool to reduce the risk of Type II errors in Quantitative Big Data—failing to discover what we could and should have seen.

Going back to the ontology, Figure 6a showed that each of the time, space and property modes had two aspects—position and intensity. Data from color video cameras etc. already contain information in all six domains. And in theory, the N-linear decomposition in Figure 7 can be used in all six domains simultaneously. But in practice, no one does that yet, to my knowledge. But we should! I think a lot of opportunities remains in this respect. Here is one suggested way forward for entrepreneurial mavericks:

3.4. The IDLE model: dual-domain modeling

In a given combination of ontological categories—for instance a series of pictures (the spatial domain), systematic variations can be compactly quantified by bilinear modeling (scores × loadings) in both abscissa (spatial position) and ordinate (intensity, e.g. light intensity for different colors). Figure 8 illustrates how I like to model a set of related images, e.g. a video sequence, in terms of what I have called the "IDLE" model. Successfully implemented, the IDLE model represents a model-based compression method that gives both high compression rates and interpretable models.

The meaning of the IDLE model name (I = D(L)+E) is explained in Figure 8. To develop a good IDLE model from video data requires that color changes and motions are estimated and modeled separately. That is not trivial (I know, for I spent about 10 years on that in the 90s, and now I am at it again). But it is doable. In 2D video describing a given scene, segmentation is needed in order to separate independent objects ("holons") to be modeled individually. It also requires motion estimation for each holon (e.g. in terms of optical flow), followed by motion compensation. Finally, the estimated motions and color changes are submitted to bilinear modeling and residual analysis.

I believe this type of multivariate dual-domain modeling, combined with elements from other sciences, will find increasing use for making sense out of the future's Quantitative Big Data. It is illustrated in the image domain in Figure 8. But it also applies

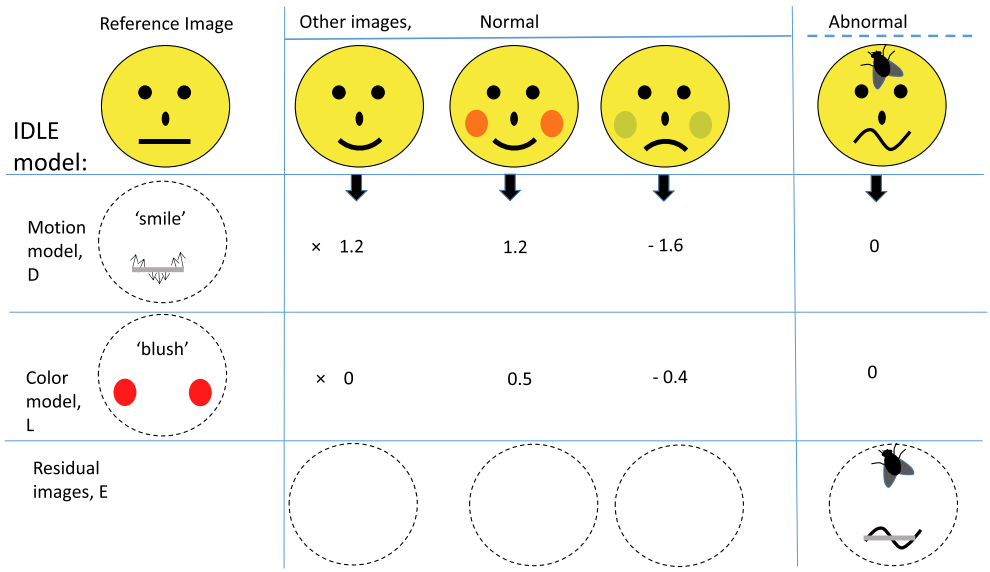


Figure 8. The dual-domain IDLE model I = D(L) + E: Intensity observed = Displacement model of (Local intensity model) + Error, illustrated in 2D: For a sequence of related images (Smiley variations, top), an idealized IDLE model (right) consists of a Reference intensity image (top), a Displacement model ("Smile loading"), an Intensity change model ("Blush model") and a set of individual Residuals (bottom row), according to the IDLE model. The Displacement model in this simple illustration consists of one single "smile" component (one soft motion pattern of the mouth) and one single "blush" component (one soft color change pattern of the cheek). Unmodeled motions, objects or colors show up in the residuals images E. When the IDLE model is applied to describe three new images (center columns), motion estimation shows that "smile" scores of 1.2, 1.2 and −1.6 allows the bland mouth expression of the reference image to be morphed into two smiles and a frown. Likewise, the "blush" scores of 0, 0.5 and −0.4 makes the second happy Smiley blush, while the third, unhappy one, get green cheeks. When later encountering an abnormal image (right column), the unexpected mouth twist and forehead fly are—ideally—left unmodeled in the Error Intensity image E.

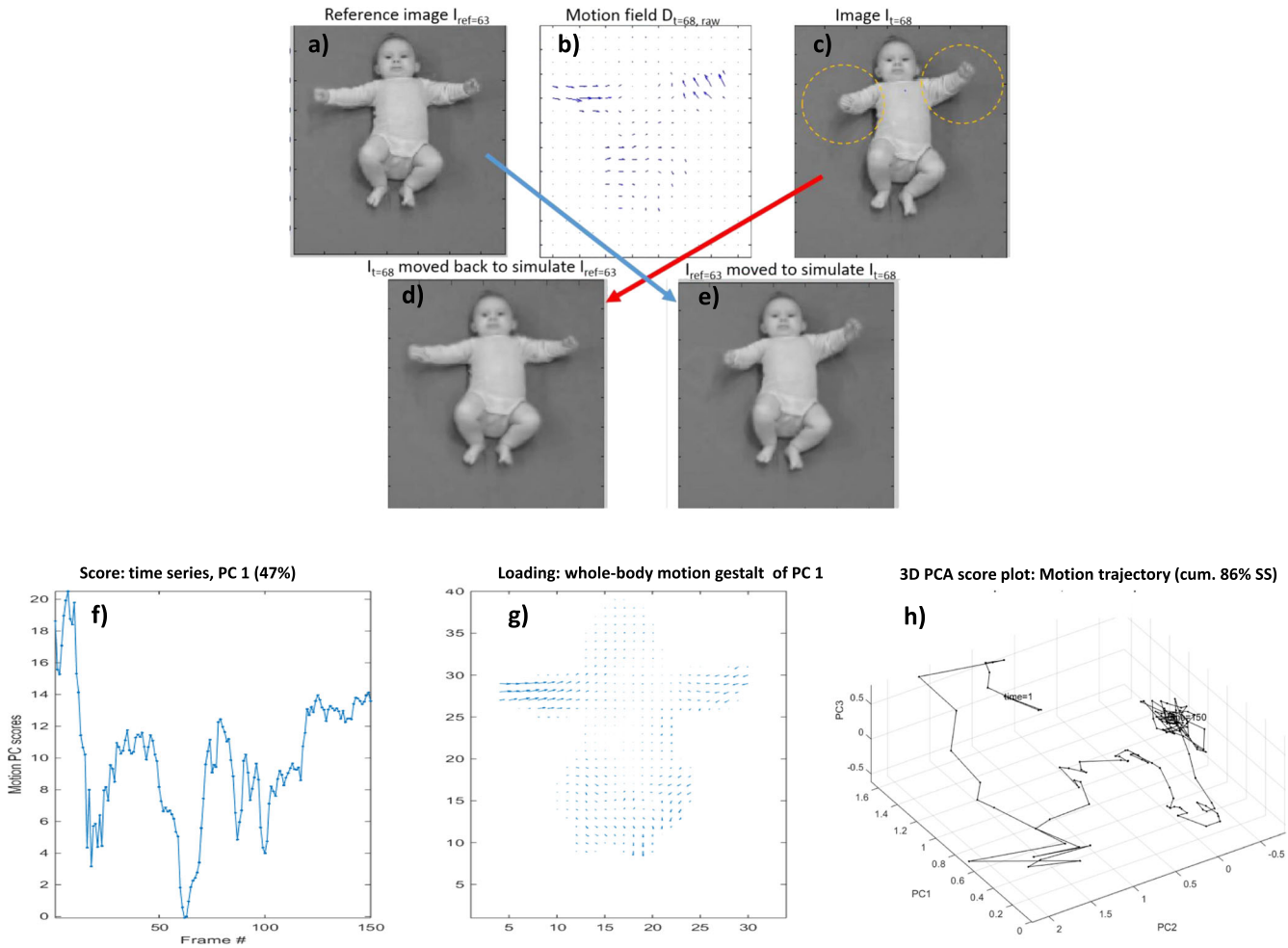


Figure 9. The IDLE principle illustrated by snap-shot from video sequence. The spontaneous motion gestalts of a baby (courtesy Lars Adde (St. Olav University Hospital and Norwegian University of Science and Technology, Trondheim, Norway). a) I_{Ref}: A chosen reference image with light intensity (frame # 63 in video) (b) $D_{n@Ref}$: Displacement field (optical flow or "smile field") showing the pattern of motions from the reference image to another image (frame n = # 68). For visual clarity, the motion arrows representing the horizontal and vertical motions, are shown for a few of the pixels only. (c) I_n: Frame n = # 68 in the same sequence. (d) $In_{@Ref}$: Frame # 68 morphed via the motion field to mimic the reference frame # 63. (e) $I_{Ref@n}$: Reference frame # 63 morphed via the motion field to mimic frame # 68. (f) A motion gestalt's time series: PCA Score of PC # 1 vs video frame #. g) A gestalt's spatial motion pattern: PCA Loading of PC # 1, folded back to 2D video pixel space. (h) 3D gestalt trajectory: Three first PCA PC score vectors, rotated to show cluster of repetitive movements.

in the time domain (time delays/phase vs. amplitude or effect shapes) and in the property domain (wavelength peak shifts vs absorbance peak heights)[¶].

Figure 9 illustrates the IDLE principle by two snap-shot from a video sequence of a baby, lent me by my colleague Lars Adde. The baby lies on its back, moving its body and head and waving its arms and legs. In the long video sequence, a representative frame (# 63) was selected as a reference, I_{Ref} (a). The motion details for another image—frame # $t = 68$, I_{68} (c) will now be shown. Comparing images (a) and (c), the arms have clearly moved, and the arm shadows too.

Figure (9b) shows the so-called motion field $D_{68@Ref}$, or just D_{68}, estimated by our proprietary implementation of optical flow estimation. The motion field D_{68} tells how each of the pixels in I_{Ref} should be moved in order to mimic the target frame I_{68} as well as possible. Moving I_{Ref} according D_{68} produced $I_{Ref@68}$ (e): The arms as well as the shadows in $I_{Ref@68}$ now resemble I_{68}. Conversely, d) shows $I_{68@Ref}$, i.e. how image I_{68} looks when it is moved back by the motion field D_{68} to mimic I_{Ref}.

To develop a two-domain motion model for a video sequence, the motion estimation and move-back operation is repeated for every frame n = 1,2,…,Ref − 1,Ref + 1,… N. (For n = IRef, $D_{Ref@Ref}$ is of course zero). Thereby, both the motion image $D_{n@Ref}$ and the moved-back intensity image $I_{n@Ref}$ (for each color channel) is represented in the *same* reference position for each image n = 1,2,…….N. $D_{n@Ref}$ and ($I_{n@Ref} - I_{Ref}$) are thus very well suited for bilinear modeling, because the meaning of each pixel is always the same for all the frames. The modelling yields bilinear spatiotemporal "smile"- and "blush" components like those illustrated in Figure 8.

[¶]I developed this dual-domain model back in the late 80s. Frank Westad and I wrote a short paper on modelling possible wavelength shifts in spectroscopy [41]; otherwise not much has been published academically on it, except in the patent literature with Jan Otto Reberg and others [42,43]. Building a business on this principle in the 90s was fun and interesting. But only till the stress almost killed me and I left it. Now, many years later, with more experience and more computer power, I am at it again with my new colleagues, and the fun is the same.

The bilinear spatiotemporal model will have temporal scores (the essential time-series) and spatial loadings (the essential motion gestalts and color change patterns). For illustration, motion fields $D_{n@Ref}$ were estimated from the reference frame (frame 63) to each of frames n = 1,2,…,125. These motion fields, with nPels pixels for the horizontal motions and nPels pixels for the vertical motions, were unfolded from their 2D representation (e.g. Figure 9b) to 1 D vectors with 2*nPels elements, generating a joint motion data table of 125 rows and 2*nPels columns. The first PCA of these motion data explained 47 % of the motion variance. This first whole-body "mathematical motion gestalt" is defined by its spatiotemporal parameters, the scores time series (Figure 9f) and its spatial whole-body motion pattern (Figure 9g).

But such raw bilinear components from PCA, PLSR etc. are not intended to be *individually* meaningful; they just provide sensible "*window frames*" for looking into the essence of the data. With 3 such PC "motion gestalts", most of the motion variance (86%) was described. The trajectory of the baby's motions in this video is summarized in the 3D "window" in Figure 9h). For interpretation, the 3D score plot has here been rotated manually to reveal a striking feature in this video—a repeated motion pattern. Other, more automatic rotation methods might have been used instead (ICA, Varimax etc.) in the scores or the loadings, to obtain so-called "simple structures" (axes corresponding to more "natural", physiological motion gestalts). More detailed interpretation will reveal which body parts are involved in which of these "natural" motion gestalts, but that is beyond the scope of this paper.

These model-based video representations may, in turn, be used for many purposes—for instance for compressed transmission, and for time series analysis and the development of more mechanistic models of ODE/PDE/Finite Element type, to map position, velocity, acceleration etc. for all pixels in the reference image, simultaneously.

I must admit that the IDLE model is my favorite: confusing to think about, somewhat tricky to implement, but versatile and powerful. I am back working on the topic after a 15-year break, and hope that IDLE modeling in the future will make it easier to handle the massive data streams from two-domain monitoring sensors for quantitative process control, improved medical diagnosis etc. So while the IDLE model itself is something to pursue also for young chemometricians, my other personal take-home message is: Never give up—even if things take more time than you like. But take a break when you need it.

3.5. Quantitative Big Data: useful and understandable

That brings me to my next favorite topic, on which I am presently working: How can our Chemometrics methods be adapted to harness the massive, overwhelming data streams that come from today's video cameras and tomorrow's multi-channel multi-way measuring devices—what I call Quantitative Big Data? Surely, buying larger and larger disks to store these overwhelming, ever-increasing data files is not an answer. Some sort of mathematical and statistical modeling is needed, to identify and extract the essence in the data.

Most multivariate data analysis methods either require all the data to be available at the time of model development, or defining a model that is modified as time goes, in a kind of "moving window". Figure 10 shows the kind of modeling that I find more in line with how we humans learn: continuously discovering new phenomena, without necessarily forgetting the past. Assume that a real-world system is continuously modified by a few input controls and then monitored by a high number of instrument outputs. This generates an ever-increasing mass of measurements: Quantitative Big Data. Alternatively, assume that a complicated mechanistic model, with a high-dimensional set of outputs, is subjected to an extensive series of computer simulations. That can also generate a cumbersome stream of data.

When the raw data stream comes from a systematic process (chemistry, physics, biology etc.) the complexity of the resulting model is often small. This means that if bi-linear modeling is used in this continuous learning process, only a limit number of components is needed. So the file size to be stored can often be drastically reduced. More importantly, the continuous soft-model learning process makes it easy for people to inspect and interpret the essence in the otherwise overwhelming data stream. The improved understanding leads to new scientific insight, and better handling of the system monitored.

Such a learning algorithm, involving not only the computer but also people, and thus bringing the human mind into the loop, is better than some of the alienating machine learning approaches employed to handle Big Data today. The way I work to develop it, non-stop massive measurement streams can thereby be compressed into their PCA-like essence without loss of valuable information. The next question is then: What is the best way to use these compact, continuous time-series of reality descriptions?

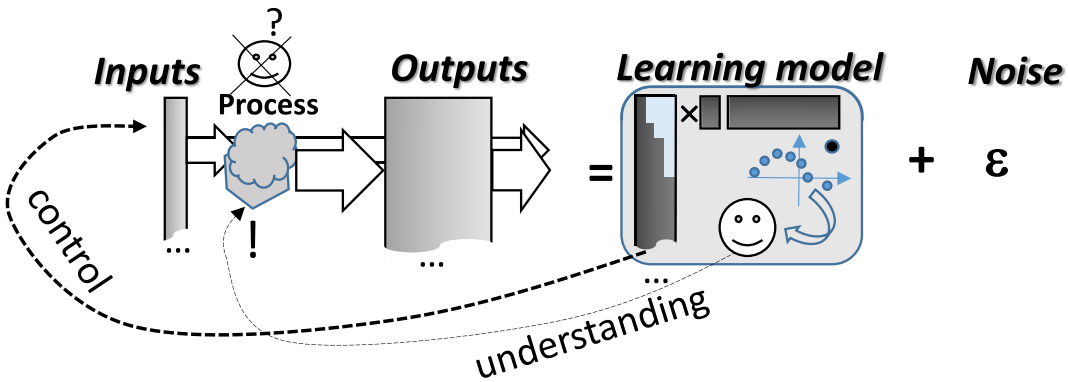


Figure 10. Making Quantitative Big Data useful and understandable. A complex, ill understood but continuous process can be a physical system continuously monitored by multichannel instrumentation, or a mathematical model subjected to extensive simulations. Such a "forever-running" system may be summarized by an automatic, self-learning approximation model. This simplified, ever-changing model should preferably of the interpretable subspace type used in Chemometrics. The process may thereby be better understood, and better controlled.

4.6. From time series data to causal insight

Having retired from my previous research positions at the Norwegian Food Research Institute and the Norwegian U. of Life Sciences at Ås, I have the privilege of now working at the Department of Engineering Cybernetics at my Alma Mater, the Norwegian U. of Science and Technology, Trondheim. There I am the only person who does not know much about PID regulators and Kalman Filters, but I see that I know more than the others about multivariate data modeling. Merged together, our fields may represent Big Data Cybernetics (Morten Breivik, pers. com.).

My question is here: How can our Chemometrics way of working contribute to process control? The first aspect of that is: How do our methods relate to dynamic modeling in terms of differential equations?

As a biochemistry student, and in my subsequent years as food scientist, I stayed away from dynamic modeling. I now see that this was because I *feared* it; seeing an integration like $\int_{-\infty}^{\infty} e^{-x^2} dx$ made me cringe. More recently, I think of integration is just a form of summation. I have found that numerical integration in, e.g. Matlab is often straight-forward. And we can often do the numerical integration once and for all and thereafter replace it by multivariate metamodeling (more later). So dynamic modeling is not particularly difficult—it is only that the notation and terminology looks so foreign.

Figure 11 will give a small illustration of how the PLS regression can be used for nonlinear dynamic modeling. By "soft" PLSR models, time series data can be converted into "harder" nonlinear dynamic models.

If we have some time series data, cross-validated Partial Least Squares Regression (PLSR) can provide interpretable, reduced-rank dynamic models. Martens et al. [28] illustrate one way in which PLSR time-series analysis can provide temporal forecasting and reveal attractor structures in multivariate dynamic systems: One or more time series vectors (Y-variables) were modeled from the same and/or additional line-shifted time series (X-variables) in a linear, reduced-rank approximation PLSR model. We also outline various cognitive computational aspects of modeling in the time domain, in light of our Chemometrics culture.

Let us first look at *linear* dynamic models. In, e.g. systems biology and control theory, simple dynamics can be expressed by linear ordinary differential equation (ODE) models of how the rate-of-change in the state of a process depends on the state itself. Data analytically, this can be written by the linear model $\mathbf{Y} \approx \mathbf{XB}$, where K time series $\mathbf{X}(N \times K) = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K]$ represent a set of K variables observed at N points in time, and rates $\mathbf{Y} = [d\mathbf{x}_1/dt, d\mathbf{x}_2/dt, \ldots, d\mathbf{x}_K/dt]$ is their temporal derivatives[‖]. Each individual column vectors $\mathbf{b}_j$ in the $K \times K$ rate constant matrix $\mathbf{B}$ in the dynamic model shows how the rate of state variable j is defined by all the K "state"-variables, including itself: $\mathbf{y}_j = \mathbf{Xb}_j$.

So we are one step closer to an apparent causal insight, because $\mathbf{Y} \approx \mathbf{XB}$ represents a set of coupled linear ODEs. I think of mechanistic ODE formulations as "how does it feel to be X inside this system? How does it work?", as opposed to the more external statistical assessment: "How does this system seem to behave?". Of course, ODEs and other mechanistic models should always be regarded with healthy skepticism, and that goes for data-driven ODEs too. But ODEs can give deeper insight into the dynamic behavior of a system:

Because the observed state variables may be expected to display natural collinearities, and because the experimental conditions seldom allow complete spanning of all combinations of all state variables, reduced-rank regression is required. Cross-validated PLSR is then an alternative for estimating the Jacobian matrix **B**. Thus, by regressing rates on states by a linear model, we can generate differential equations from time series data. And using a reduced-rank linear regression method, like PLSR with cross-validation to estimate the optimal rank and jack-knifing to estimate the precision of the resulting Jacobian, we should be able to make reliable ODE's even from highly intercorrelated time series, by standard chemometric methodology.

Then, going beyond standard chemometrics, we can understand the dynamic behavior of the system by analyzing the mathematical properties of the obtained Jacobian matrix **B**. That is outside the comfort zone of most chemometricians, including mine. But control theory can then tell us how the system will behave wrt short- and long-term temporal stability, primarily from its eigenvalue properties**.

In general, I believe that the IDLE-like extensions of bilinear modeling in time, space and properties × position change and intensity change may be seen as powerful multivariate generalizations of the bi-linear part of a Taylor expansion. But sometimes these linear (or quadratic) constraints limit us too much. Therefore, we chemometricians should look more into the generalizations of useful techniques I have learnt from Psychometrics (I had my early exposure to professional multivariate data modeling among psychometricians in the 70s and early 80s, mainly in the US).

Figure 11 shows how to look for the right *nonlinear* differential equation model that have caused a certain observed set of time series in a system. It employs standard chemometric PLSR, but on X-variables represented at what in Measurement Theory is called the NOMINAL level:

A given hypothetical system was characterized by three state variables, observed as functions of time. (I received these four error-free three-dimensional time series, generated by computer simulations, i.e. numerical integration from four different sets of initial states without being told the underlying causal structure of the ODE used for generating them by simulation). I computed

[‖]The ODE structure is here written in standard chemometric/statistical regression notation. In control theory the notation is different.

[**]In Chemometrics we might be inclined to perform singular value decomposition of **B**, since that bilinear decomposition would correspond to our understanding of a model with reduced rank A ($A \leq K$): The "latent states" are defined from the "manifest states": $\mathbf{T}_A = \mathbf{XV}_A$, and the rates from these score vectors $\mathbf{T}_A$: $\mathbf{Y} \approx \mathbf{T}_A\mathbf{Q}_A'$. This means that the obtained Jacobian is bilinear: $\mathbf{B}_A = \mathbf{V}_A\mathbf{Q}_A'$. This is described in more detail in [28]*.

If so desired, we may reformulate how we span the A-dimensional subspace, by a transformation matrix $\mathbf{C}_A$ ($A \times A$), e.g. by ICA, MCR, Varimax rotation etc. $\mathbf{D}_A = \mathbf{XV}_A\mathbf{C}_A$, $\mathbf{Y} \approx \mathbf{D}_A\mathbf{C}_A^{-1}\mathbf{Q}_A'$. But unless we go beyond this linear model formulation, we get $\mathbf{B}_A = \mathbf{V}_A\mathbf{C}_A\mathbf{C}_A^{-1}\mathbf{Q}_A' = \mathbf{V}_A\mathbf{Q}_A'$. So I don't think that would affect the Jacobian.

How should we then analyze the properties of the quadratic matrix $\mathbf{B}_A$(K x K)? Singular value decomposition (svd) of $\mathbf{B}_A$ would only give us back the X- and Y-subspaces of $\mathbf{V}_A$ and $\mathbf{Q}_A'$, which we already know from the PLSR. And eigenvalue decomposition of $\mathbf{B}_A'\mathbf{B}_A$ or $\mathbf{B}_A\mathbf{B}_A'$, which are both quadratic and *symmetric*, would of course also have given us the same basis vectors as svd of $\mathbf{B}_A$. However, the in control theory, Jacobian $\mathbf{B}_A$ itself is submitted to eigenvalue decomposition. And $\mathbf{B}_A$ is *not symmetric*. Therefore—depending on the values in $\mathbf{B}_A$—its eigen-analysis may result in some negative, and even *complex* eigenvalues.

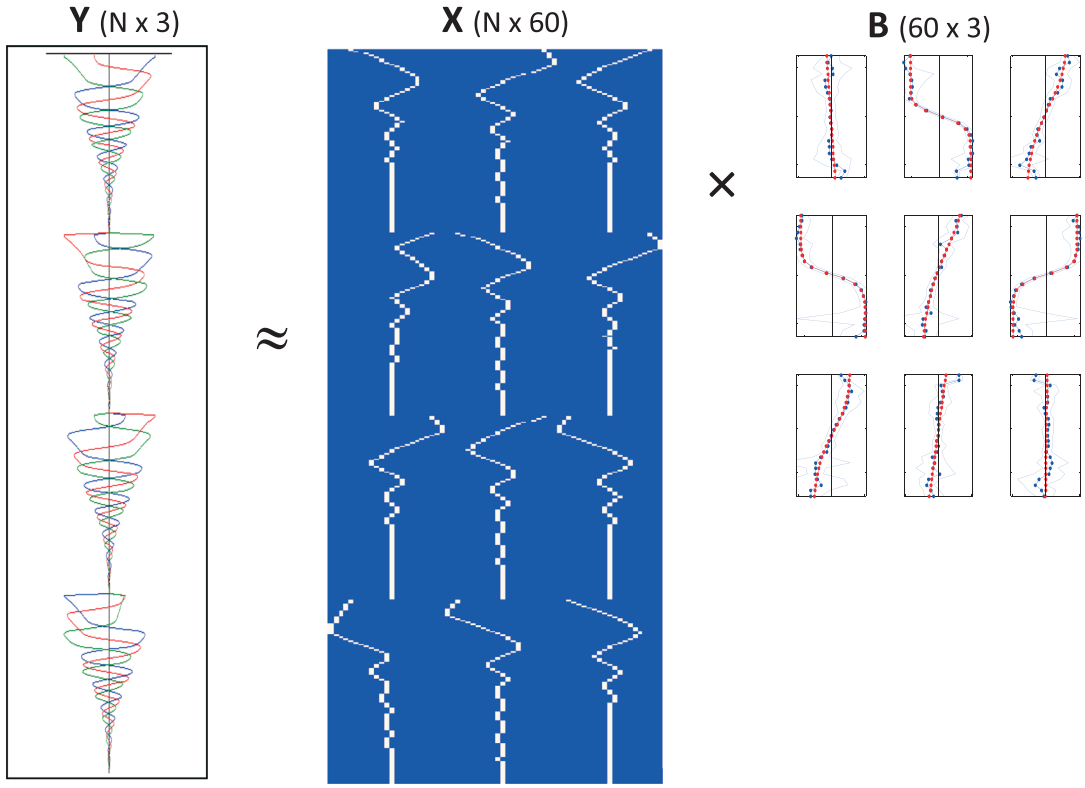


Figure 11. ODE development by Nominal-level PLSR: dynamics and nonlinearity. Data-driven generation of nonlinear differential equation (ODE) system. Example of nonlinear (nominal-level) PLS regression Adapted from Martens et al. 2013 [28]. Adapted from Martens et al. 2013 [28] A complex system is here to be characterized in terms of a nonlinear dynamic mathematical model, linking three state variables. *Middle*: The input data consisted of four sets of time series, each containing them time series of the three different state variables obtained by numerical integration for a new set of initial states. Each of the three state variables were split into 20 category variables (white = 1, dark = 0) and the set of these 60 nominal variables were together used as X-variables. Left: The three state variables were also differentiated with respect to time, and used as three Y-variables. Conventional PLS regression was employed based on the linear model of rates = *f* (states) (i.e. Y ≈ XB). Cross-validation showed that four PCs gave optimal prediction of rates Y from statcategories X. *Right*: The nominal-level regression coefficients B at optimal PLS regression rank was finally split to show how different levels of each of the tree states affected each of the three rates.

the time derivatives of each of these time series, and defined them as 3 Y-variables. I could have used the three time series directly as X-variables, and obtained the Jacobian as the regression coefficient **B** in $\mathbf{Y} = \mathbf{XB} + \mathbf{F}$. But that did not work here; I found that the causal dynamics behind the data must be highly nonlinear.

Therefore I chose to split each of the observed quantitative time series into 20 nominal (category) variables, with values 0 or 1 as illustrated in Figure 11, and defined these 60 indicator variables as my X-variables. The nonlinear "Jacobian" was the estimated as the regression coefficient at optimal rank, by a sampling-balanced, cross-validated/jackknifed PLSR. The results showed negative, linear "cis"-effects proportional (degradation effects of each state variable on itself). For 6 of the "trans"-effects, three of them were found to be close to 0 and three were found to be highly nonlinear sigmoids. When I sent these results back to my colleague who had generated the data, he confirmed that I had indeed found the causal structure behind the data.

This is just one little demo example, using one of the nonlinear techniques from psychometric Measurement Theory, nominal scaling. Others (ordinal scaling) are also available, in addition to the Interval (like Farenheit or Centigrade temperature scales) and Ratio scaling (Kelvin temperature scale).

I advise students of chemometrics to combine our standard tools from chemometrics, with methods from other disciplines, like image analysis and control theory. Armed with such a method arsenal, a surprisingly wide range of data can be analyzed. For instance, the next section shows how Chemometrics can pay something back to our friends in mathematics:

4.7. Multivariate metamodeling: models of models

The subspace methods that we often use in Chemometrics have proven useful for getting a better grip on main-stream mathematical modeling, particularly when the models are large and/or intricate, and thus slow to compute and difficult to overview. By massive statistically designed computer simulations with such models, followed by subspace analysis of the large resulting Input and Output data tables, several different benefits may be obtained. In [6] we reviewed the use of conventional PLSR, N-way PLSR and non-linear PLSR for revealing the actual behavior of a wide range of mathematical model types:

Highly reduced experimental designs, like Latin Hyper-Cube design or Optimized Binary Replacement design allow models with many inputs (e.g. 10 or 20 input parameters) to be probed systematically at many levels in factorial designs, without experiencing combinatorial explosion.

"Classical" metamodeling in the causal direction, describing the model outputs in terms of the defined model inputs, gives

both sensitivity analysis and model overview, and can give substantial computational speed-up.

"Inverse" metamodeling, reversing the modeling direction to describe the inputs from the outputs, can substantially speed up the fitting of complicated mathematical models to experimental data. Because the simulations and the metamodeling can be done once and for all, problems with long, iterative search processes and getting stuck in local minima are avoided.

But because many mechanistic models are mathematically sloppy, i.e. that several input parameter combinations give more or less the same model output [31], such inverse modeling can sometimes be difficult. Then it is better to compare the experimental data directly to the raw simulation data. We have called that the Direct Look-Up method Isaeva et al. (2012a) [32] and (2012b) [34]—which is a simple version of so-called Case-based Reasoning.

The ambiguity associated with sloppy models can changed from being a nuisance and source of confusion to a source of deeper insight by chemometric analysis [33]: The so-called neutral parameter set, obtained when fitting a sloppy model to a given set of empirical data, represents a sample of "all" the parameter combinations that give satisfactory fit to these empirical data. PCA or PLSR of this neutral parameter set shows the "structure of doubt" in the modeling process, in terms of linear subspaces or non-linear sub-manifolds of equivalent parameter settings.

Figure 12 illustrates another use of multivariate metamodeling: Faster and more comprehensive fitting of nonlinear models to data [34]. In this case it concerns how to parameterize growth curves. We needed a way to describe the complex dynamics of a spatially heterogeneous system (2D electrophoresis gels in proteomics). Sarin Nhek therefore monitored the color developments with a video camera, each time obtaining more than 100 000 different measured "growth curves" of pixel darkness z_1 as functions of time x_1, $z_{1,meas.} = f(x_1)$; one for each pixel.

To quantify these curves, we wanted to fit each of them to a nonlinear growth model—the logistic function, with two parameters p_1 and p_2: $z_0 = \frac{1}{1+e^{(-p_1 x_0 + p_2)}}$ in a suitable time range. Initial attempts to do that by conventional, iterative curve-fitting failed miserably—many of the optimizations ended up in useless local minima, and the whole procedure was prohibitively slow and had to be terminated.

So instead, we ran the logistic function, for a selected range of abscissa values x_0, with several thousand different value combinations for parameters p_1 and p_2 chosen according to a factorial design. The black curves in Figure 12a) show some of them, after having resized both axes in simulations $z_0 = f(x_0)$ to range [0,1] at 100 different abscissa values, forming a set of several thousand simulation curves $z_{sim.} = f(x)$. Likewise, the measured curves $z_{1,meas.} = f(x_1)$ were resized and interpolated to form observed curves in the same range, $z_{meas.} = f(x)$.

Based on these simulations, two fast, quantitative predictors could then be developed:

(1) Multivariate calibration: A so-called inverse multivariate metamodel, whereby parameters p_1 and p_2 were described as two regressand variables Y from the resized response vector z_{sim} (100 regressor variables X) by cross-validated PLSR. Once established, the unknown parameters p_1 and p_2 were for each pixel simply predicted by replacing z_{sim} by z_{meas}.
(2) Case-based reasoning: A Direct Look-Up, whereby each measured curve z_{meas} was compared to the simulated curves z_{sim}, and all plausible matches listed, possibly with a simple local linear interpolation.

Again, the former metamodeling is extremely fast, but requires the model to have unique one-to-one relationships between input parameters and output curves. The latter one may be slightly slower, but works also for "sloppy" models with non-unique input/output relationships, which gives a more informative overview of the ambiguity in the parameterization for the chosen model, the logistic function.

However, there is also ambiguity in the choice of the mechanistic model itself. A number of alternative mathematical models could have been chosen. How to compare and interconnect these mechanistic mathematical models? So we tried to find "the mathematics behind the mathematics" of simple line-curvature:

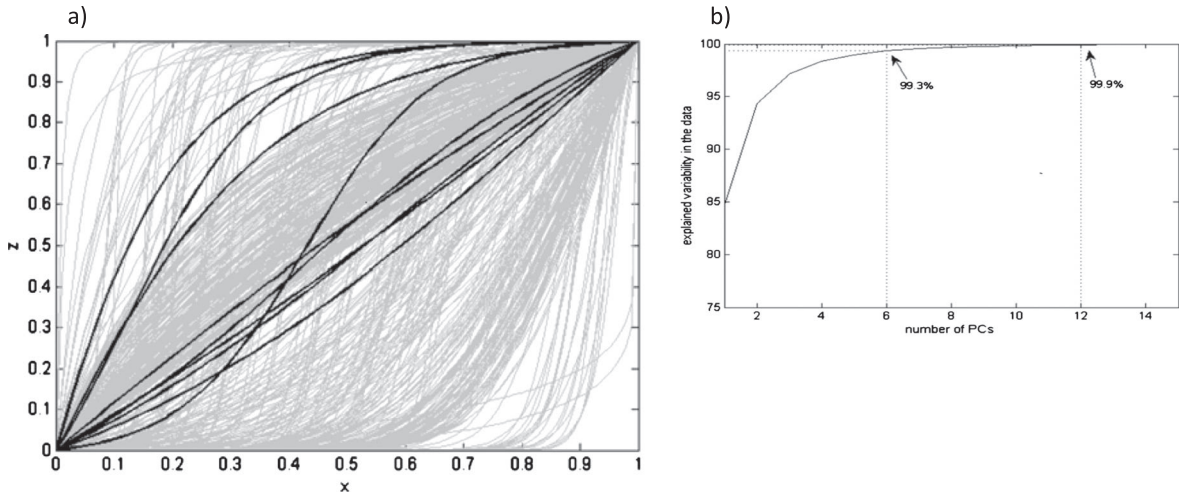


Figure 12. Example of multivariate metamodeling: The modelome of line curvature. (a) Thirty-eight widely different mathematical models of line curvature, z = f(x, parameters), represented by i = 1,2,…, 500 out of N > 17 000 simulations using different parameter combinations (Isaeva et al. 2012a, [32]), after normalization of x and z to [0,1]. Dark lines: one of the 38 models, the Logistic Function. (b) PCA of all N curves. Percent variance vs # of principal components, showing that the modelome of 38 different mathematical models of line curvature can be inter-linked quantitatively with high precision (99.9% correctly explained variance) via a 12-dimensional bilinear metamodel.

Different fields of science were scrutinized, and thirty-eight relevant, but widely different mathematical models of line curvature were found, each capable of producing smooth sigmoids and archs. Among the 38 models, we included the Hill function, the two-parameter logistic function, the five-parameter logistic function, the Gompert function, various kinetic functions, trigonometric functions, cumulative statistical distribution functions and ODE-integrals. All of them were nonlinear and thus difficult to fit to curve data. For each of the 38 alternative nonlinear dynamic models, massive computer simulations were done, applying different parameters and initial value combinations according to statistical designs, generating the outputs $\mathbf{z}_{0,\ sim} = f(\mathbf{x}_0)$. In total, more than 17 000 simulations were performed. For each of the obtained curves, both the input vector $\mathbf{x}_0$ and output vector $\mathbf{z}_0$ were normalized to [0,1], yielding $\mathbf{z}_{sim} = f(\mathbf{x})$. Figure 12a) shows 500 out of the $N > 17\,000$ normalized curve simulations.

A joint PCA was performed on these $N > 17\,000$ normalized curves. Figure 12b) shows that 95% of the variance among them was explained by only two PCs: >99% by 6 PCs and about 99.9% by 12 PCs[††].

Summarizing, the 38 competing curvature models had so diverse mathematical form that it was very difficult to compare them directly. But they are easily compared via their joint PCA metamodel, linking their behavioral repertoires of output curves. This behavior can, in turn, be linked to their input parameter values. Thus, for any given parameter combination of any of the curvature models, we can check if any of the 37 other models behave similarly, and if so, identify their parameter values.

We consider our collection of the 38 curvature models, with the data base and joint PCA metamodel, as a first estimate of the "*modelome* of line curvature". Technically, this use of the multivariate soft modeling of line curvature turned out to be less unique than we thought at the time. It is similar to a technique in computer vision, the Kanade–Lucas–Tomasi (KLT) feature tracker [35–37], which also relies on PCA-like analysis of lots of features. Moreover, the obtained loadings from the joint PCA model in our specific case—smooth growth curves—looked very similar to cosine functions with different frequencies. So a Fast Fourier Transform (FFT) or a Discrete Cosine Transform (DCT) might have been used instead, in which the orthonormal loadings are preselected instead of generated by PCA of the data. But our general idea of merging many competing mechanistic models into one joint metamodel is perhaps new?

In the future, similar modelomes may be established for other classes of models with comparable behavior. For instance, it might perhaps be possible to make modelome-libraries of different classes of linear or nonlinear ODEs and PDEs, thus doing away with the need for time-consuming numerical integration.

4.8. Combining human and technical data

The twin fields of Chemometrics and Sensometrics share many aspects with each other and with, e.g. Psychometrics. For one thing, we use factor analytical subspace-methods like PCA and PLSR for open-ended, transparent data-driven modeling suitable for human interpretation. Moreover, this methodology makes it easy to combine "immaterial" human response data with "material" technical data.

For instance, from the modelome-of-line-curvature data (Figure 12), Isaeva et al. (2012a) [32] submitted print-outs from a representative subset of the thousands of simulated curves to quantitative sensory descriptive analysis, using a sensory panel from food science. Once the panelists had developed a suitable vocabulary and used it to profile the selected curves, a PLS regression model was developed to describe sensory profile Y from the curves' joint metamodel scores X. Using this PLSR model, it was now possible to predict how humans would have verbally described each of the >17 000 curves—if they had seen them, or vice versa.

A similar sensory profiling was used to describe complex behavior of a nonlinear spatiotemporal model of cell differentiation [38]. Using their senses and language capabilities, the human assessors were able to reveal a new and totally unexpected mathematical pattern type.

To handle the future deluge of data, artificial intelligence will hardly be intelligent enough. Human interpretation will be required. Future chemometricians should learn to respect Sensory Science as a wonderful way to give meaning to measurements and models.

4. A TWO-WAY BRIDGE ACROSS THE MATH GAP

Working in science concerns having a job at all, and having fun on the job. But not just that: Scientists are talented people, and I think we should use our talents for something meaningful. My own university has the motto "Knowledge for a better world". I subscribe to that. Referring to Figure 5, it is clear that we are all somewhat different, and should use our talents differently. Some should bake the cake, some should sell it and some should develop new cakes.

Figure 2 outlined the difference between the inductive and the deductive mind-set in science. With that in mind, let us now return to my original concern in Figure 1, the problematic gap between the mathematical sciences and non-mathematical sciences. I sincerely believe that our chemometrics culture and tool-box, together with other soft-modeling disciplines such as sensometrics, psychometrics and morphometrics, and other pragmatic, real-world oriented fields such as image analysis, signal processing and control theory plus realistic statistical validation [44], can bridge that gap. If we do that, we have done science and society a favor.

Figure 13 concerns how to build a bridge across the math gap. The bridge is two-way, from the inductive to the deductive mode of working, and vice versa. Assume that a given, real-world system is studied by experimental observation by some scientist. Assume also that other scientists study the same system theoretically, by mechanistic mathematical modeling. Traditionally, these two groups of scientists would not meet (Figure 1). However, both groups would control their system by variations in certain inputs and obtaining certain outputs. If both groups used multivariate soft modeling to link their inputs and outputs, they could develop a fruitful two-way communication bridge, with several benefits:

Variables a) and b) in Figure 13 represent the inputs and outputs common to the real experiments and the computer simulations. To the extent that the mathematical model describes the

[††]In addition, there are offset- and slope parameters for the normalization of x and z.

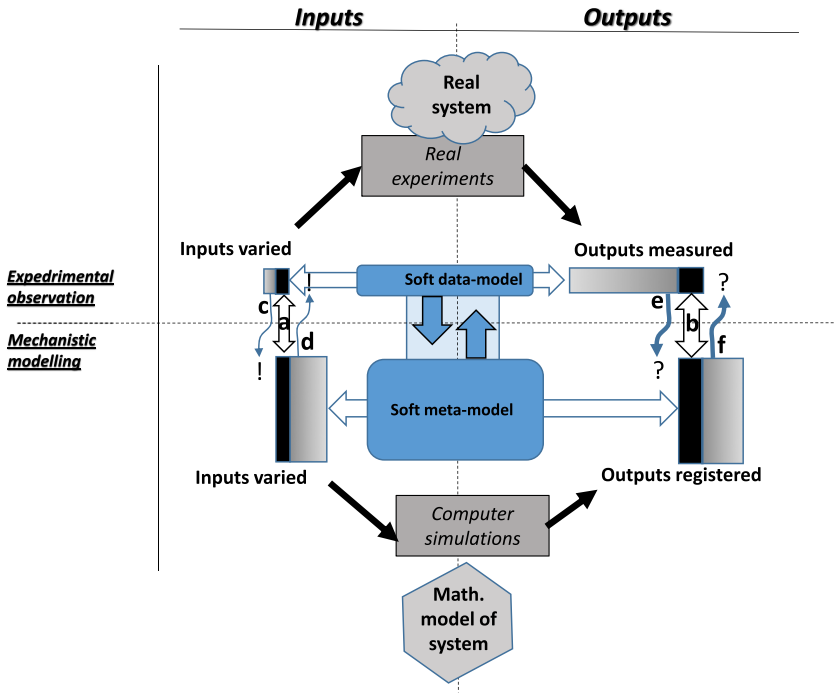


Figure 13. How to combine data-driven and theory-driven research. (a) Input parameters common to experiment and model simulations, to be compared. (b) Output parameters common to experiment and model simulations, to be compared. (c) Experiment input parameters to add to improved model. (d) Model input parameters to add to next experiment. (E) Experiment output variables to add to improved model. (F) Model output variables to add to next experiment.

system well, the soft-modeling should show the same behavior patterns for these common inputs and outputs. If not, their discrepancies should either lead to critical assessment of the real experiments, or used for developing data-driven model extensions, as explained by Martens [39].

Variables c) and d) represent inputs involved only in the real experiment or only in the simulation experiment. Among these, variables shown to be important for predicting the outputs, should be included and varied consciously in future simulations and/or experiments.

Likewise, variables e) and f) represent outputs involved only in the real experiment or only in the simulation experiment. Among these, variables shown to be important for predicting the outputs should be monitored, for improvements of future experiments and/or simulations.

I admit, I have not yet had the chance to demonstrate this diplomatic bridge-building in practice.

A lot of good science has been done and is still done by performing small, critical exercises, be it hypothesis-based experiments or critical simulation attempts. But that is a slow and risky process. I am convinced that in the future, science will make much more use of high-throughput, high-dimensional measuring instruments as well as massive, well-designed computer simulations and multivariate metamodeling. If that prediction comes true, then I believe our two-way bridge across the math gap is bound to build itself, rendering both "looking-good" statistics and "macho" modeling obsolete.

But that will require a major change in the way math and statistics is taught to the students in chemistry, biology, medicine and applied sciences. A communal, international effort is called for to develop course contents and teaching styles suited for students of all major types of personality types.

5. CONCLUSIONS

To handle the future's Quantitative Big Data, we need a lot more applied mathematical modeling, statistical assessments and graphical interpretation tools. I have here outlined my view of how Chemometrics culture in the future may contribute to building a two-way bridge across the math gap in science and technology, in particular wrt the biomedical field:

On one hand, the data-driven multivariate soft modeling—in the tradition of conventional Chemometrics—can convey to the theoretically oriented mathematical modeling community how the Real World seems to behave, empirically. On the other hand, multivariate metamodeling, based on methods, e.g. from Chemometrics, can convey to more math-averse empiricists that mathematical modeling of, e.g. non-linear dynamic mechanisms is not as difficult as they may think.

Acknowledgements

It is difficult to define how ideas arise and develop. Much of what I presented builds on impulses gathered from others, over many years. I am indebted to so many. In particular, Finn Tschudi, Forrest Young, Joseph Kruskal, Magni Martens, Lars Munck, Svante Wold, Bruce Kowalski, Edmund Malinowski, Arne Tenningen, Svein Berg, Tormod Næs, Frank Westad, Solve Sæbø, Rolf Ergon, Beata Walczak, Jan Otto Reberg, Klaus Diepold, Stig Omholt, Erik Plahte, Peter Hunter, Dalibor Stys, Helge Brovold, Øyvind Stavdahl and Ole Morten Aamo have been important for my understanding as outlined here, and I thank them deeply. Jan Otto Reberg and Nina Heilemann are thanked for constructive comments, and Cyril Rochebusch is thanked for patient editorial prodding.

Lars Adde and Helge Brovold are thanked for generously sharing their data for this paper.

REFERENCES

1. Martens H. On the calibration of a multivariate instrument for quantitative estimation of individual components in a mixture. In *Proceedings, Nordic Symposium on Applied Statistics*, Høskuldsson A, Conradsen K, Sloth Jensen B, Esbensen K (eds.). Institute of Mathematical Modelling: Technical University of Denmark: Lyngby, Denmark 1980.
2. Martens H. *Multivariate Calibration—Quantitative Interpretation of Non-Selective Chemical Data*. Dr. Technol. thesis. 1985; Norwegian U. of Science and Technology, Norway.
3. Martens H. Reliable and relevant modelling of real world data: a personal account of the development of PLS regression. *Chemom. Intell. Lab. Syst.* 2001; **58**: 85–95.
4. Martens H, Næs T. *Multivariate Calibration*, J. Wiley & Sons Ltd: Chichester UK, 1989.
5. Martens H, Martens M. *Multivariate Analysis of Quality. An Introduction*. J. Wiley & Sols Ltd Chichester UK, 2001.
6. Tøndel K, Martens H. Analyzing complex mathematical model behavior by PLSR-based multivariate metamodeling. *WIREs Computational Statistics* 1994; **6**: 440–475, DOI: 10.1002/wics.1325 −600.
7. Brovold H. Four ways into mathematics (in Norwegian). PhD Thesis, Dept. Psychology, Norwegian U. of Science and Technology, Norway 2014; (http://www.realfagsrekruttering.no/wp-content/uploads/2013/10/Brovold-2013-4-veier-inn-i-matematikken-.pdf).
8. Ziliak ST, McCloskey DN. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives (Economics, Cognition, and Society). 2008; Michigan Publishing.
9. Lakoff G, Núñez R. *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*. 2000; Basic Books. ISBN 0-465-03771-2.
10. Núñez R, Lakoff G. The cognitive foundations of mathematics: the role of conceptual metaphor. In *Handbook of Mathematical Cognition*, Campbell J (ed.). Psychology Press, New York, 2014 ISBN-10: 1138006068.
11. Kahneman D. *Thinking Fast and Slow*, Farrar, Straus and Giroux: New York, 2011.
12. Martens H, Martens M. NIR spectroscopy—applied philosophy. In *Near Infra-Red Spectroscopy. Bridging the Gap between Data Analysis and NIR Applications*, Hildrum KI, Isaksson T, Naes T, Tandberg A (eds.). Ellis Horwood: Chichester UK, 1992 1–10.
13. Bakeev KA, Esbensen KH, Paasch-Mortensen P Process sampling: theory of sampling—the missing link in process analytical technologies (PAT). In: *Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries*, Bakeev K A (ed.). (2nd edn). Wiley: Chichester UK, 2010; ISBN: 978-0-470-72207-7.
14. Martens H, Martens M. Modified jack-knife estimation of parameter uncertainty in bilinear modelling (PLSR). *Food Quality and Preference* 2000; **11**(1–2): 5–16.
15. Sæbø S, Almøy T, Aarøe J, Aastveit AH. ST-PLS: a multi-directional nearest shrunken centroid type classifier via PLS. *J. Chemometrics* 2008; **22**(1): 54–62.
16. Indahl UG, Liland KH, Næs T. *J of Chemometrics* 2009; **23**: 495–504.
17. Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *R Stat Soc Series B Stat Methodol* 2010; **72**(1): 3–25.
18. Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* 2012; **118**: 62–69.
19. Liland KH, Martens H, Sæbø S. Distribution based truncation for variable selection in subspace methods for multivariate regression. *Chemom. Intell. Lab. Syst.* 2013; **122**: 103–111.
20. Westad F, Marini F. Validation of chemometric models—a tutorial. Analytica Chimica Acta 2015; *in press*.
21. Rubingh CM, Martens H, van der Voet H, Smilde AK. The costs of complex model optimization. *Chemom. Intell. Lab. Syst.* 2013; **125**: 139–146.
22. Anderssen E, Dyrstad K, Westad F, Martens H. Reducing overoptimism in variable selection by cross-model validation. *Chemom. Intell. Lab. Syst.* 2006; **84**: 1–2.
23. Bro R. PARAFAC. *Tutorial & applications. Chemom. Intell. Lab. Syst* 1997; **38**: 149–171.
24. Westerhuis JA, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemometrics* 1998; **12**: 301–321.
25. Hassani S, Martens H, Qannari EM, Hanafi M, Kohler A. Model validation and error estimation in multi-block partial least squares regression. *Chemom. Intell. Lab. Syst.* 2012; **117**: 42–53.
26. Biancolillo A, Måge I, Næs T. Combining SO-PLS and linear discriminant analysis for multi-block classification. *Chemom. Intell. Lab. Syst.* 2015; **141**: 58–67.
27. Martens H. Domino PLS: A framework for multi-directional path modelling. Proc. PLS'05 Intl Symposium "PLS and related methods". (Eds. Aluja, T., Casanovas, J., Vinzi, V.E., Morineau, A., Tenenhaus, M.) SPAD Groupe Test&Go), 2005; 125–132.
28. Martens H, Tøndel K, Tafintseva V, Kohler A, Plahte E, Vik JO, Gjuvsland AB, Omholt SW. PLS-based multivariate metamodeling of dynamic systems. In: *New Perspectives in Partial Least Squares and Related Methods. Springer Proceedings in Mathematics & Statistics* 2013; **56**: 3–30. DOI: 10.1007/978-1-4614-8283-3_1.
29. Gadamer H-G, Hermeneutics and social science. *Philosophy Social Criticism/Cultural Hermeneutics 1975* 1975; **2** (4): 307–316, D. Reidel Publishing Company, Dordrecht, Holland.
30. Martens H, Høy M, Wise BM, Bro R, Brockhoff PMB. Pre-whitening of data by covariance-weighted pre-processing. *J. Chemometrics* 2003; **17**: 153–165.
31. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP, universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* 2007; **3**(10): e189. doi:10.1371/journal.pcbi.0030189.
32. Isaeva J, Martens M, Sæbø S, Wyller JA, Martens H. The modelome of line curvature: many nonlinear models approximated by a single bi-linear metamodel with verbal profiling. *Physica D: Nonlinear Phenomena* 2012a; **241**: 877–889.
33. Tafintseva V, Tøndel K, Ponosov A, Martens H. Global structure of sloppiness in a nonlinear model. *J. Chemometrics* 2014; **28**: 645–655. DOI: 10.1002/cem.2651.
34. Isaeva J, Sæbø S, Wyller JA, Nhek S, Martens H. Fast and comprehensive estimation and fitting of complex mathematical models to massive amounts of empirical data. *Chemom. Intell. Lab. Syst* 2012b; **117**: 13–21.
35. Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision. *International Joint Conference on Artificial Intelligence* 1981. https://www.ri.cmu.edu/pub_files/pub3/lucas_bruce_d_1981_1/lucas_bruce_d_1981_1.pdf 674–679.
36. Tomasi C, Kanade T. Detection and Tracking of Point Features. Carnegie Mellon University technical report CMU-CS-91-132, April 1991.
37. Shi J, Tomasi C. Good features to track. IEEE Conference on Computer Vision and Pattern Recognition 1994; **593**.
38. Martens H, Veflingstad SR, Plahte E, Martens M, Bertrand B, Omholt SW. The genotype–phenotype relationship in multicellular pattern-generating models—the neglected role of pattern descriptors. *BMC Syst. Biol.* 2009; **3**: 87doi:10.1186/1752-0509-3-87.
39. Martens H. The informative converse paradox: windows into the unknown. *Chemom. Intell. Lab. Syst.* 2011; **107**: 124–138.
40. Lowe EJ. *The Four-Category Ontology: A Metaphysical Foundation for Natural Science*. Oxford University Press: 2005; ISBN 0199254397.
41. Westad F, Martens H. Shift and intensity modeling in spectroscopy—general concept and applications. *Chemom. Intell. Lab. Syst.* 1999a; **45**: 361–370.
42. Martens H, Reberg JO. Method and apparatus for decoding video images. *US Patent* 1995; **6**: 046,773.
43. Martens H, Reberg JO. Method and apparatus for coordination of motion determination of multiple frames. 1996; *US PATENT* # 6,157,766.
44. Westad F, Marini F. Validation of chemometric models - a tutorial. *Chemom. Intell. Lab. Syst.* 2015; in press.