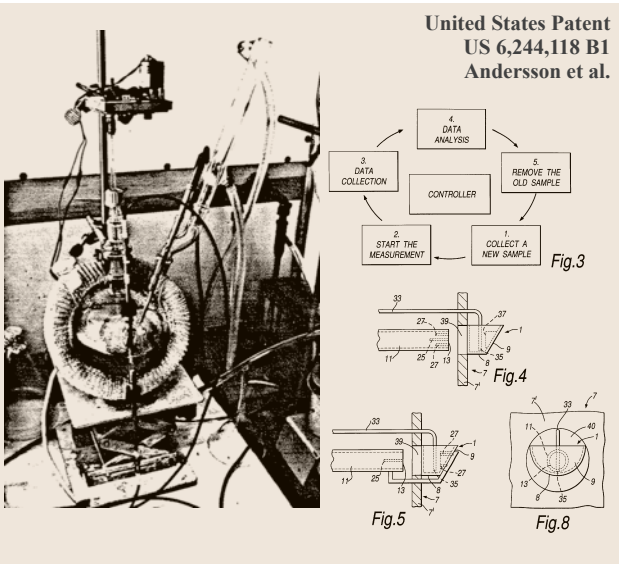
Behind the 2010 Kowalski Prize

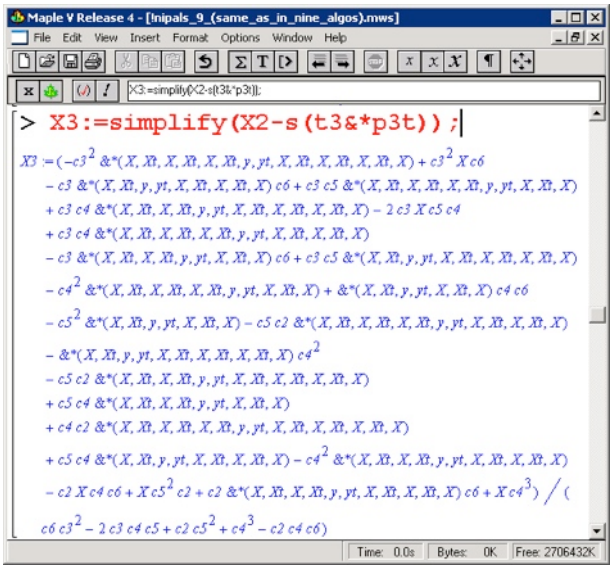
I was asked to write about the story behind the 2010 Kowalski Prize in Chemometrics. I will need a few of pages to do so.

The story began in 1997. At that time, I was hired by AstraZeneca to carry out research on in-line near infrared (NIR) process analysis, and I worked on a PhD thesis on this theme at Lund University. I had been interested by partial least squares (PLS) regression ever since 1993 when I completed my Master's thesis on on-line NIR using a Guided Wave instrument from Tecator. I don't know why, but I was very attracted to the PLS from when I first saw it. I wanted to know more, but the information that I was interested in was not to be found anywhere, so I began my own investigations. First, I worked with paper and pen to carry out the mathematics, but the expressions became huge. It was impossible to do all the algebra correctly, so I decided to write a package for Maple, which is software for symbolic mathematics. With the Maple package, I could be sure that I obtained the correct expressions. The problem was then that I did not realize how they could be used, so I stored the files in a folder on my computer and left matters there. I thought that they may be useful later.

I met all the stars within the field of chemometrics early in my career at a conference in Reykjavik in 1994. It is unlikely that I will ever attend a better conference; it was a fantastic group and an amazing environment. There was good chemistry between me and three young men from what was then a new group studying chemometrics at KVL in Copenhagen. Their leader was Lars Munck, whom I soon found to be one of the most inspiring people that it is possible to ever meet. I kept in touch with the group and went to them to share and discuss ideas in an enjoyable manner. They helped me with some early three-way applications within pharmaceutical analysis and I discussed an idea I had on variable selection with Rasmus Bro and Lars Nørgaard. They gave me confidence and understood the potential of my idea for variable selection and encouraged me to publish it. I wrote a short communication on my work and sent it to the *Journal of Chemometrics* in 1998. I felt that similar ideas would appear soon, which turned out to be correct, so I wanted my work published as soon as possible. For this reason, the quality of the submission was perhaps not as good as it could have been. The comments from the editor-in-chief were that I should "use formulas and equations to present theory. No Matlab commands". I knew this would be difficult because there was no explicit mathematical expression available for the PLS regression vector. A year later, I realized that I could use my work on the symbolic mathematics of the original PLS algorithm, NIPALS. You could say that the rejection of my paper on variable selection prompted me to work on the theory. The first step would be to determine an explicit expression for the PLS regression vector. I began the task knowing that it would take years.



The interest in chemometrics started with in-line NIR for a Master Thesis 1993 and PhD Thesis 1999.



Maple was used for algebraic calculations used to better understand the theory of PLS.

After the AstraZeneca project, I joined FOSS to do image analysis. The calibration work was ANN based, the PLS research continued and kept me awake at nights.

In 1999, I joined FOSS to work with image analysis and artificial neural networks. This was one of the topics that had appealed to me the most during my time at Lund University/AstraZeneca. The new job at FOSS was exciting and I got to travel around the world. I did not work much with PLS in my daily work, but I heard some interesting news—that researchers at the Chemometrics Laboratories at FOSS in Hillerød had been using my variable selection method for quite some time. I was delighted to learn that it had been suggested to them by Lars Munck's group. The method was being used even though it hadn't been published! Even better, it was being used in the company that I was working for! At this time, all the work on the theory of the PLS regression vector was done at night, during flights, or whenever I had a chance, and progress was very slow. One of the projects I worked on as part of sales and support of image analysis solutions within grain quality inspection was a new type of analyzer to be used in Japan. Because of this project, I went more and more often to Japan, and one day, the General Manager at FOSS in Japan asked me to move to Japan to work. I moved in 2004 and soon met my wife, but that is another story.

My most important work from a revenue point of view was (and still is) to work with new customers, introduce new instruments, or do both simultaneously. These are important tasks for all application specialists at FOSS. However, we cannot work on new projects all the time, and in my case, I was also involved with customer training, internal training, and the support of everyday sales and service activities. I also dedicated some of my office hours to my own research. Someone once asked me: "But are you allowed to do that? You are not hired to do that, are you?" My answer was that I have never asked if I was allowed or not. I said that it is something that I have to do, and that this was accepted by FOSS Japan from the beginning. In this way, I was able to continue my research, sometimes in the office, sometimes in the backseat of a car on the way to yet another customer, sometimes at airports, and sometimes at business hotels. My main work was to push the quality of calibrations and datasets, and thus, I could not carry out full-time research. In any event, because ideas take time to grow, I felt that I did not need more time. Instead, I thought it was good to keep in touch with the "real" chemometrics working with customers data as much as possible. It also forced me to find new ways of doing research, such as in immensely packed trains while commuting to and from the office 90 minutes each way every day. In such a train, you can't even read a book because the person next to you is standing too close to you; I could only choose between listening to my iPod or closing my eyes and trying to do the mathematics. I'm sure that I was the only one in the train thinking about PLS (with a smile on my face!).

I finally discovered how the PLS regression vector could be expressed explicitly and termed it Krylov-PLS. Later it would turn out that I had discovered the worst way of doing PLS. I installed LaTeX on a couple of computers so that I could write the mathematics properly, and I wrote down my findings. I had been working on this alone for such a long time, and I had been away from the scientific community in the field of chemometrics, so I preferred to have someone to write with. I sent my paper to Rasmus Bro and Lars Nørgaard. Although they were very positive, they did not have time to be coauthors. Instead they gave me a list of the best researchers in the field to work with. The list of names was surprisingly short, which confirmed to me for the first time that my scoop was at a good level and somewhat unique. I started with Sijmen de Jong, a very famous chemometrician from the Unilever Research Laboratories in the

Chemometrics and image analysis took me to Japan where the PLS research continued.

I got one more reason to continue my research in Japan, but that is another story.

Netherlands, because I had met him in Iceland in 1994 and because I was impressed by his publications. He gave me very valuable feedback on the formulas and equations regarding typing errors, and he advised me how to write things; for example, the Krylov sequence is normally written in the opposite way to a polynomial. He thought the content was too limited to publish a new paper and said, "It will not be easy to publish these results as they are." He was right. For some reason, I lost contact with Sijmen. Maybe he was tired of PLS after a whole career with it, or maybe he just thought that I might as well do it myself. In any event, I am very grateful for his valuable comments. I realized that the content wasn't substantial enough and that I would have to do something drastic to change the theme of the paper. I added another PLS algorithm, something that had come up in a research side-track from my time in Japan. I later called it direct-scores PLS. Perhaps I could make a comparison of the direct-scores PLS and the explicit expression for the regression vector to be able to publish a paper? I had noticed that the two methods had very different numerical properties even though I knew they were theoretically identical, but I did not yet realize that this would be something that I could write about.

I had only heard of Rolf Manne from the chemometrics group at the University of Bergen once before, when I went to Umetrics in Umeå where Svante Wold mentioned Rolf Manne in a comment

$$\mathbf{b}_A = \begin{bmatrix} \mathbf{X}'\mathbf{y} & \mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{y} & \dots \end{bmatrix} \begin{bmatrix} \mathbf{X}\mathbf{X}'\mathbf{y} & \mathbf{X}\mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{y} & \dots \end{bmatrix}^{+} \mathbf{y}$$

In 2007, I finally found the explicit expression for the PLS regression vector. Later it turned out that I had also found the worst way of doing PLS regression.

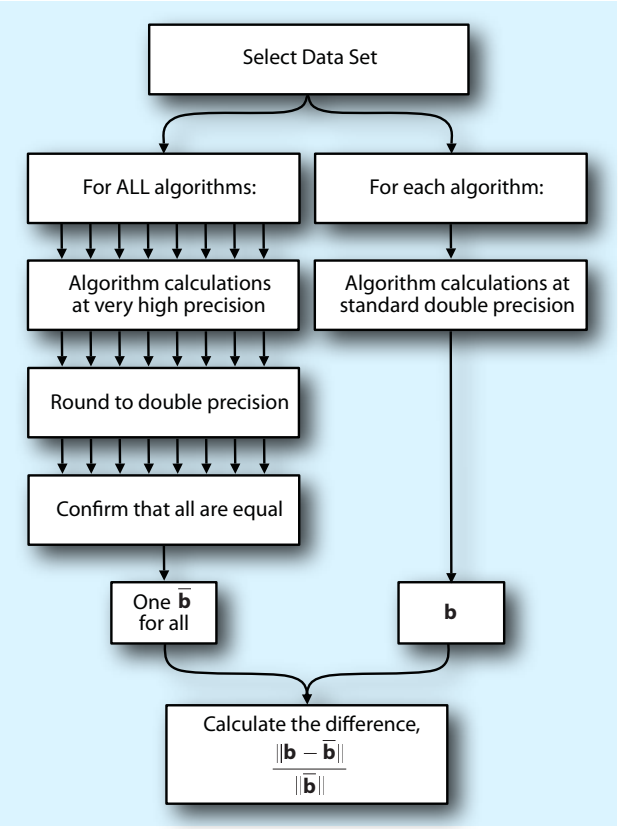
on the deflation of Y-variables (that it was not theoretically necessary to do it, but that the Umetrics software did it anyway). I wondered if it would be a good idea to ask Rolf Manne for comment. I found out via Google that he was a professor in theoretical chemistry and you never know about those guys. In any event, I wrote to him to ask for his comments and it turned out to be one of the best decisions I ever made. He was very helpful, but he was not interested in becoming a coauthor. Although we are yet to meet in person, it turned out that we had more to talk about than PLS. We have become friends through our correspondence and I hope to meet him some day. However, the words of Sijmen still echoed in my head. Would it be possible to make my work substantial enough for publication?

In the summer of 2008, I received an email from Rolf Manne telling me that there was a big discussion on the International Chemometrics Society listserver, an Internet discussion forum for chemometricians. I still remember checking the discussion very late one night in a hotel in the city of Naha in Okinawa, southern Japan. It seemed to be more or less a flame war! The debate was over a paper for which Rolf Manne was a coauthor, and it discussed model residuals and "consistency" in PLS regression. I still remember how I was shocked at what I read and immediately, I became very interested. As soon as possible, I made sure that I had all the scientific papers necessary, and I started reading to make up my own opinion on the matter. I then realized that the debate on which algorithm should be preferred and the residuals was the greatest thing that could happen to me. Suddenly, I had a context in which my ideas on PLS could be presented and discussed and would be interesting to other chemometricians or people

In 2008 there was a fight going on in the chemometrics community. The war on residuals was the best thing that could happen to me. Suddenly I had a context to put my results into that would make it interesting to many readers.

within applied mathematics. I then decided to add other algorithms to the paper and to compare them. The war on residuals was the final key to writing a paper that clearly explained the differences. I thought that the missing concept in the war on PLS residuals was the point about numerical differences between algorithms. Although the discussion on residuals regarded a theoretical difference, it was not clear why the algorithms gave different results for the cases where they theoretically should have been the same. While starting to write the paper, Rolf Manne was my mentor and his most helpful critique was that I should not use NIPALS as a reference, as I had done originally. This guided me toward the idea of comparing against the theoretically exact regression vector.

But how would I be able to compare the algorithms? The algorithms all gave slightly different results. I then got the idea to use more decimals in the calculations. I downloaded and installed the Multiple Precision Toolbox for Matlab written by Ben Barrowes and hacked it a little to get it working to my requirements. I found that when using more than 500 decimals in the calculation, and after rounding to normal double precision (which is what I used in most software packages), all algorithms gave identical results for up to 40 factors, a very high number, definitely more than enough for

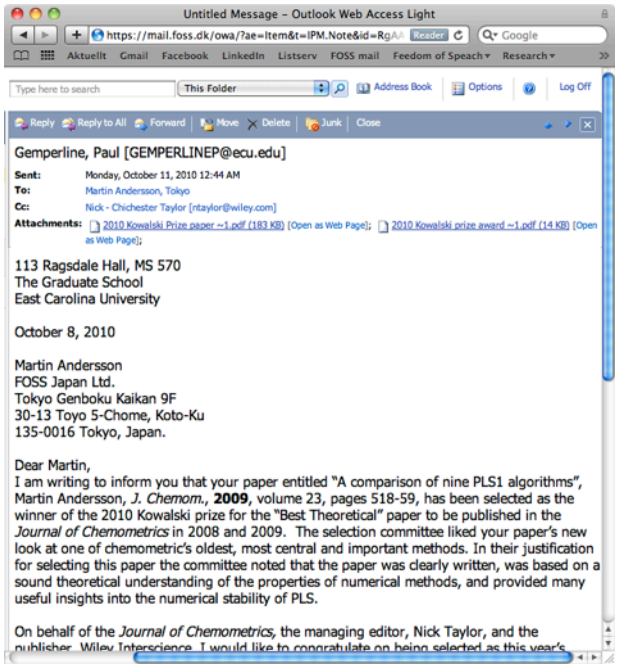


Workflows for high-precision calculations and standard double-precision calculations.

all four real datasets that I used. The only problem was that the calculations took days or weeks to complete. I remember that I added a function that would automatically email me when new results had been produced or when an error appeared. In this way, I could do other things without nervously checking the results all the time. After completing the high-precision calculations, I could compare each algorithm with something that was as close to exact as possible instead of referring to NIPALS. For a while, I was very absorbed in this new approach to the problem and I worked very hard to get the paper in good shape. I very rarely saw my family, but I knew that this project would soon come to an end. One day I said to Rolf Manne, "now I will submit". After 11 years of work, even if Rolf thought that more could have been done before submitting, I sent the paper to the *Journal of Chemometrics*. Anyway, I thought the reviewers would have many comments.

The handling of the review by the *Journal of Chemometrics* was good. I have to admit that responding to the three peer reviewers meant that I was busy once again. Never before have any of my papers been reviewed by three professors. Perhaps the journal had selected three professors because the paper had high potential? I thought so, and I was encouraged even if responding to their comments was a lot of work. The comments that I had received before from Sijmen de Jong and Rolf Manne were valuable, but the comments from the reviewers were even better, and I regard the reviewers as coauthors. Most of their comments greatly improved the paper and only a few redundant tables made it worse in my opinion. The problem was really how to make all the reviewers happy because they had different opinions. The heaviest work resulting from the comments was to add a table on the calculation speeds for huge datasets such as those of metabonomics data having more than 100,000 variables. I only had 32-bit Windows on my computers at work, and it could not handle such large datasets. The only solution was to install a 64-bit Linux operating system. I came to realize that Ubuntu Linux is now in very good shape and that Octave can be used as an alternative to Matlab. Thank you, reviewers, whoever you are!

After a year and three rounds of peer review, the paper was accepted for publication. I was of course very happy, but then it was time to get the typesetting right. This was also a challenge, but the editor-in-chief of the *Journal of Chemometrics* continually supported me. He agreed that making things easer for the readers was more important than making things easier for the typesetters. Finally, I could accept the typesetting for publication



I got a mail: "You are the winner." It sounded like spam, but it turned out to be better.

and it was published. after 12 years of work I was very happy to receive many emails commenting on my work from around the world. It was also interesting to see that Barry Wise and his Eigenvector Research, Inc. implemented my direct-scores PLS only a few months after its publication—an impressively quick time for implementation. They even took it further and added the PLS2 version, to my great delight. At the end of summer 2010, Barry Wise posted some comments on the International Chemometrics Society listserver about my paper, and a quite lively discussion among experts from all around the world started. It was nice for me to see my work being discussed, and what a pleasure it was for me to drop a comment or two!

On October 6th 2010, I received an unusual email. In fact, it looked like spam in the form of "you are the winner..." When I opened the email, I saw that it was about my paper and I had been awarded the Kowalski Prize—the finest award in chemometrics for the best theoretical paper published in the last two years. What a surprise! I didn't even know that I had been nominated. At first I thought a prize like this was not important, but then I realized that because of the award, I would be able to explain to people who couldn't understand the mathematics that I had produced something of worth. Furthermore, I could be sure that the top tier of international researchers understood the significance of what I had presented. This made me very happy. I felt that I could be understood and I even came to realize that I myself was in the top tier! What a feeling! I think that the award is not only important to me, but also to my employer, FOSS. It clearly states that we are at the highest international level, even regarding details like the theoretical aspects of chemometrics. I am just one of the many application specialists in the FOSS organization, and I know that all of us are doing a very good job, and that we are essential to our solutions. The award confirms to our sales staff and our customers that they can be confident that the people behind our solutions always strive for and can obtain world-class results.

This story has a very pleasant ending. Just before I received the award, I was notified that the chemometrics section at FOSS will be managed by an old friend from the beginning of my career in PLS research. Lars Nørgaard has just joined FOSS to lead the chemometrics team. In my upcoming work in PLS research, I hope that the variable selection method will someday be published. Afterward, I would like to present a completely new idea on how to perform PLS regression, which is unlike any previously presented PLS method. After that I might retire.

2010-10-09
Martin Andersson, FOSS Japan K.K.