Preface

In psychology, *regression* means degeneration, or return to a prior, lower level of development. And, indeed, speaking of statistical methods sounds somewhat outdated today — all those neural networks and fuzzy systems being now available.

However, it has been recognized that the new soft computing methods are no panacea. These methods cannot find models any better than the more conventional ones if there is not enough information available in the data to start with. On the other hand, many celebrated soft computing applications could have been solved with age-old methods — assuming that somebody were familiar with them. After all, neural networks can only operate on the statistical properties visible in the data. Why not concentrate on these statistical properties directly?

The starting point here is that *when analysing complex systems, simple methods are needed.* When the system under study is very complicated, one needs data analysis methods that are reliable and fast and that give possibility for closer analysis.

Unfortunately, the statistical literature seems to be mathematically rather unpenetrable for normal engineers (for those that are not too ashamed to admit that). Researchers seem to represent the results in such sophisticated forms that the very simple and elegant ideas remain hidden — the problem is that the powerful methods may not be applied in practice. What is more, different methods have been developed in different research communities: It is difficult to see the connections between the methods when the approaches and notations differ. New methods are constantly being developed; there exists no one-volume book that would cover the whole field in sufficient detail.

This text tries to show how simple the basic ideas are and how closely related different methods turn out to be. The approach is rather pragmatic, many proofs being omitted for readability. All the methods are presented in a homogeneous framework. It is crucial that a student recognizes that all formulas and algorithms are based on simple underlying principles; everything can be questioned and nothing needs to be believed as some kind of unpenetrable "secret wisdom". One objective is to show that there is still room for new ideas and innovations. As the reader can see, there exist plenty of ideas waiting to be explored and exploited — indeeed, one is very near to frontier science, maybe able to cross the boundary towards new discoveries (examples of such exploratory experiments are indicated by stars "*").

The theoretical methods are supported by Matlab routines. The implemented "vanilla" algorithms are by no means optimized; their main purpose is also to show that the methods are by no means unpenetrable.

In addition to the printed version, this report is available in public domain in PostScript format. The Matlab Toolbox and the textual material can be accessed through the HTML page at the Internet address

`http://saato014.hut.fi/hyotyniemi/publications/01_reportXXX.htm.`

The earlier version of this text, with the name "Multivariate Regression — Techniques and Tools" was published in 2001. Despite its various shortcomings, it received a positive acceptance. The printed version was "sold out" a long time ago. This interest was the motivation to try and fix some of the holes that were left open.

In the previous version, applications were not discussed: Now there is the last chapter that tries to present examples of not so evident but potential ways of applying the new tools. Second, the fields of physical first-principles modeling and data-oriented modeling are not so distinct that no connections could be found; the new appendices concentrates on this kind of discussions, fitting the structural considerations into the domain of numeric manipulations in a more or less seamless way.

National Technology Agency of Finland (TEKES) has provided funding during the research under several project frames, and this support is gratefully acknowledged.

Heikki Hyötyniemi

List of symbols

The same variable names are used consistently in the theoretical discussion and in the accompanying `Regression Toolbox` for `Matlab`, if possible.

- A, B, C, D: Matrices determining a state-space system
- c: Arbitrary constant, scalar or vector
- i, j: Matrix and vector indices
- e, E: Measurement error vector and matrix, dimensions $m \times 1$ and $k \times m$, respectively
- ϵ: State error, dimension $n \times 1$
- f, F: Vector and matrix defining a linear mapping
- $g(\cdot)$: Any function (scalar or vector-valued)
- γ, Γ: Constraint vector and matrix, respectively
- $J(\cdot)$: Cost criterion
- I, I_n: Identity matrix (of dimension $n \times n$)
- k: Time index, sample number (given in parentheses)
- K: Kalman gain
- m: Dimension of the output space
- M: Arbitrary matrix (or vector)
- n: Dimension of the input space / Non-compressed feature space
- N: Dimension of the latent space / Number of levels or substructures
- P: Error covariance matrix, dimension $d \times d$
- $p(\cdot)$: Probability density
- R: Covariance matrix (or, more generally, *association matrix*)
- u, U: Input vector and matrix, dimensions $\nu \times 1$ and $k \times \nu$, respectively
- v, V: Combined data vector and matrix (x and y together), dimensions $m + n \times 1$ and $k \times m + n$, respectively
- ν: Unprocessed measurement vector / Dimension of the vector u in dynamic systems
- w, W: Weight vector and matrix, respectively
- x, X: Data vector and matrix, dimensions $n \times 1$ and $k \times n$, respectively. In the case of a dynamic system, state vector and matrix, dimensions $d \times 1$ and $k \times d$, respectively.

- y, Y: Output data vector and matrix, dimensions $m \times 1$ and $k \times m$, respectively
- z, Z: Latent data vector and matrix, dimensions $N \times 1$ and $k \times N$, respectively
- ξ, ζ: Arbitrary scalars
- λ, Λ: Vector of eigenvalues and eigenvalue matrix, respectively
- μ, η: Lagrange multipliers
- θ, ϕ: Reduced base, dimensions $n \times N$ and $m \times N$, respectively
- Θ, Φ: Matrices of data basis vectors, dimensions $n \times n$ and $m \times m$, respectively

Notations

- $\mathbf{M}$: Unprocessed data
- $\bar{M}$: Mean value of M (columnwise mean matrix if M is matrix)
- $\hat{M}$: Estimate of M
- $\tilde{M}$: Error in M / Erroneous M
- M': M modified (somehow)
- M^i: Power of M / Level i data structure
- M_i: Column i for matrix M / Element i for vector-form M
- M^T: Transpose of M
- $M^\dagger$: Psedoinverse of M (for definition, see page 21)
- $\mathrm{E}\{M\}$: Expectation value of M
- M_{test}, M_{est}: Independent testing data or run-time data
- $\left(\begin{array}{c|c} M_1 & M_2 \end{array} \right)$: Partitioning of a matrix
- $\underset{\xi \times \zeta}{M}$: Matrix dimensions ($\xi$ rows, ζ columns)

Abbreviations

- CCA/CCR: Canonical Correlation Analysis/Regression (page 102)
- CR: Continuum Regression (page 98)
- CA: Cluster Analysis (page 205)
- CLR: Constrained Linear Regression (page **??**)

- DA or FDA: (Fisher) Discriminant Analysis (page 208)
- EIV: Error In Variables model (page 68)
- FOBI: Fourth-Order Blind Identification (page 115)
- GHA: Generalized Hebbian Algorithm (page 138)
- ICA/ICR: Independent Component Analysis/Regression (page 109)
- MLR: Multi-Linear Regression (page 63)
- NNR: Neural Networks based Regression (page 129)
- OLS: Orthogonal Least Squareas (page 72)
- PCA/PCR: Principal Component Analysis/Regression (page 77)
- PLS: Partial Least Squares regression (page 95)
- RR: Ridge Regression (page 73)
- SOM: Self-Organizing Map (page 126)
- SPC: Statistical Process Control (page 90)
- SSI: SubSpace Identification (page 143)
- TLS: Total Least Squares (page 185)

Contents

[1]The simulations were carried out by Mr. Hans-Christian Pfisterer
[2]The simulations were carried out by Mr. Olli Haavisto, M.Sc.

A Good Book Must Have Obscure Quotes

A theory has only the alternative of being wrong. A model has a third possibility — it might be right but irrelevant.

— M. Eigen

Statistics in the hands of an engineer are like a lamppost to a drunk — they're used more for support than illumination.

— B. Sangster

Like other occult techniques of divination, the statistical method has a private jargon deliberately contrived to obscure its methods from non-practitioners.

— G. O. Ashley

Make the model as simple as possible — but not simpler!

— A. Einstein

There are three kinds of lies: lies, damned lies, and statistics.

— B. Disraeli

> In earlier times, they had no statistics, and so they had to fall back on lies. — Stephen Leacock

Torture the data long enough and they will confess to anything.

— unknown

> If at first it doesn't fit, fit, fit again.
>
> — J. McPhee

Data! Data! Data! I can't make bricks without clay!

— S. Holmes

...

Why's the bell-shaped curve called normal?
Is it normal to be so formal?
There's nothing mean about the mean.
Its just average, as is clearly seen.
And what's so standard about that deviation?
It's a really malicious creation.
Confusing students is its only function.
It frustrates and mystifies, in conjunction.
...

— *"On statistical terminology"* by C. Lation