Partial least squares regression

From Wikipedia, the free encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
Español
Français
Edit links

Partial least squares regression (PLS regression) is a statistical method that bears some relation to principal components regression; instead of finding hyperplanes of minimum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Because both the *X* and *Y* data are projected to new spaces, the PLS family of methods are known as bilinear factor models. Partial least squares Discriminant Analysis (PLS-DA) is a variant used when the Y is categorical.

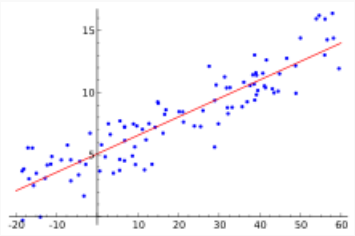
PLS is used to find the fundamental relations between two matrices (*X* and *Y*), i.e. a latent variable approach to modeling the covariance structures in these two spaces. A PLS model will try to find the multidimensional direction in the *X* space that explains the maximum multidimensional variance direction in the *Y* space. PLS regression is particularly suited when the matrix of predictors has more variables than observations, and when there is multicollinearity among *X* values. By contrast, standard regression will fail in these cases (unless it is regularized).

The PLS algorithm is employed in partial least squares path modeling,[1][2] a method of modeling a "causal" network of latent variables (causes cannot be determined without experimental or quasi-experimental methods, but one typically bases a latent variable model on the prior theoretical assumption that latent variables cause manifestations in their measured indicators). This technique is a form of structural equation modeling, distinguished from the classical method by being component-based rather than covariance-based.[3]

Partial least squares was introduced by the Swedish statistician Herman Wold, who then developed it with his son, Svante Wold. An alternative term for PLS (and more correct according to Svante Wold[4]) is *projection to latent structures*, but the term *partial least squares* is still dominant in many areas. Although the original applications were in the social sciences, PLS regression is today most widely used in chemometrics and related areas. It is also used in bioinformatics, sensometrics, neuroscience and anthropology. In contrast, PLS path modeling is most often used in social sciences, econometrics, marketing and strategic management.

Part of a series on Statistics

Regression analysis



Models

Linear regression · Simple regression · Ordinary least squares · Polynomial regression · General linear model

Generalized linear model · Discrete choice · Logistic regression · Multinomial logit · Mixed logit · Probit · Multinomial probit · Ordered logit · Ordered probit · Poisson

Multilevel model · Fixed effects · Random effects · Mixed model

Nonlinear regression · Nonparametric · Semiparametric · Robust · Quantile · Isotonic · Principal components · Least angle · Local · Segmented

Errors-in-variables

Estimation

Least squares · Ordinary least squares · Linear (math) · **Partial** · Total · Generalized · Weighted · Non-linear · Non-negative · Iteratively reweighted · Ridge regression

Least absolute deviations · Bayesian · Bayesian multivariate

Background

Regression model validation · Mean and predicted response · Errors and residuals · Goodness of fit · Studentized residual · Gauss–Markov theorem

Statistics portal

v · t · e

Contents [hide]

Underlying model [edit]

The general underlying model of multivariate PLS is

$$X = TP^\top + E$$
$$Y = UQ^\top + F$$

where X is an $n \times m$ matrix of predictors, Y is an $n \times p$ matrix of responses; T and U are $n \times l$ matrices that are, respectively, projections of X (the *X score*, *component* or *factor* matrix) and projections of Y (the *Y scores*); P and Q are, respectively, $m \times l$ and $p \times l$ orthogonal *loading* matrices; and matrices E and F are the error terms, assumed to be independent and identically distributed random normal variables. The decompositions of X and Y are made so as to maximise the covariance between T and U.

Algorithms [edit]

A number of variants of PLS exist for estimating the factor and loading matrices T, U, P and Q. Most of them construct estimates of the linear regression between X and Y as $Y = X\tilde{B} + \tilde{B}_0$. Some PLS algorithms are only appropriate for the case where Y is a column vector, while others deal with the general case of a matrix Y. Algorithms also differ on whether they estimate the factor matrix T as an orthogonal, an orthonormal matrix or not.[5][6][7][8][9][10] The final prediction will be the same for all these varieties of PLS, but the components will differ.

PLS1 [edit]

PLS1 is a widely used algorithm appropriate for the vector Y case. It estimates T as an orthonormal matrix. In pseudocode it is expressed below (capital letters are matrices, lower case letters are vectors if they are superscripted and scalars if they are subscripted):

```
function PLS1(X, y, l)
X^(0) ← X
w^(0) ← X^T y / ||X^T y||, an initial estimate of w.
t^(0) ← X w^(0)
for k = 0 to l
    t_k ← t^(k)^T t^(k)  (note this is a scalar)
    t^(k) ← t^(k) / t_k
    p^(k) ← X^(k)^T t^(k)
    q_k ← y^T t^(k)  (note this is a scalar)
    if q_k = 0
        l ← k, break the for loop
    if k < l
        X^(k+1) ← X^(k) − t_k t^(k) p^(k)^T
        w^(k+1) ← X^(k+1)^T y
        t^(k+1) ← X^(k+1) w^(k+1)
end for
define W to be the matrix with columns w^(0), w^(1), ..., w^(l−1).
Do the same to form the P matrix and q vector.
B ← W (P^T W)^−1 q
B_0 ← q_0 − P^(0)^T B
return B, B_0
```

This form of the algorithm does not require centering of the input X and Y, as this is performed implicitly by the algorithm. This algorithm features 'deflation' of the matrix X (subtraction of $t_k t^{(k)} p^{(k)^T}$), but deflation of the vector y is not performed, as it is not necessary (it can be proved that deflating y yields the same results as not deflating.). The user-supplied variable l is the limit on the number of latent factors in the regression; if it equals the rank of the matrix X, the algorithm will yield the least squares regression estimates for B and B_0

Extensions [edit]

In 2002 a new method was published called orthogonal projections to latent structures (OPLS). In OPLS, continuous variable data is separated into predictive and uncorrelated information. This leads to improved diagnostics, as well as more easily interpreted visualization. However, these changes only improve the interpretability, not the predictivity, of the PLS models.[11] L-PLS extends PLS regression to 3 connected data blocks.[12] Similarly, OPLS-DA (Discriminant Analysis) may be applied when working with discrete variables, as in classification and biomarker studies.

Software implementation [edit]

Most major statistical software packages offer PLS regression.[*citation needed*] The 'pls' package in R provides a range of algorithms.[13]

See also [edit]

- Feature extraction
- Data mining
- Machine learning
- Regression analysis
- Canonical correlation
- Deming regression
- Multilinear subspace learning
- Principal component analysis
- Total sum of squares

Further reading [edit]

- Kramer, R. (1998). *Chemometric Techniques for Quantitative Analysis*. Marcel-Dekker. ISBN 0-8247-0198-4.
- Frank, Ildiko E.; Friedman, Jerome H. (1993). "A Statistical View of Some Chemometrics Regression Tools". *Technometrics* **35** (2): 109–148. doi:10.1080/00401706.1993.10485033.
- Haenlein, Michael; Kaplan, Andreas M. (2004). "A Beginner's Guide to Partial Least Squares Analysis". *Understanding Statistics* **3** (4): 283–297. doi:10.1207/s15328031us0304_4.
- Henseler, Joerg; Fassott, Georg (2005). "Testing Moderating Effects in PLS Path Models. An Illustration of Available Procedures".
- Lingjærde, Ole-Christian; Christophersen, Nils (2000). "Shrinkage Structure of Partial Least Squares". *Scandinavian Journal of Statistics* **27** (3): 459–473. doi:10.1111/1467-9469.00201.
- Tenenhaus, Michel (1998). *La Régression PLS: Théorie et Pratique. Paris: Technip*.
- Rosipal, Roman; Kramer, Nicole (2006). "Overview and Recent Advances in Partial Least Squares, in Subspace, Latent Structure and Feature Selection Techniques". pp. 34–51.
- Helland, Inge S. (1990). "PLS regression and statistical models". *Scandinavian Journal of Statistics* **17** (2): 97–114. JSTOR 4616159.
- Wold, Herman (1966). "Estimation of principal components and related models by iterative least squares". In Krishnaiaah, P.R. *Multivariate Analysis*. New York: Academic Press. pp. 391–420.
- Wold, Herman (1981). *The fix-point approach to interdependent systems*. Amsterdam: North Holland.
- Wold, Herman (1985). "Partial least squares". In Kotz, Samuel; Johnson, Norman L. *Encyclopedia of statistical sciences* **6**. New York: Wiley. pp. 581–591.
- Wold, Svante; Ruhe, Axel; Wold, Herman; Dunn, W.J. (1984). "The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses". *SIAM Journal on Scientific and Statistical Computing* **5** (3): 735–743. doi:10.1137/0905052.
- Garthwaite, Paul H. (1994). "An Interpretation of Partial Least Squares". *Journal of the American Statistical Association* **89** (425): 122–7. doi:10.1080/01621459.1994.10476452. JSTOR 2291207.
- Wang, H., ed. (2010). *Handbook of Partial Least Squares*. ISBN 978-3-540-32825-4.
- Stone, M.; Brooks, R.J. (1990). "Continuum Regression: Cross-Validated Sequentially Constructed Prediction embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression". *Journal of the Royal Statistical Society, Series B* **52** (2): 237–269. JSTOR 2345437.
- Wan Mohamad Asyraf Bin Wan Afthanorhan. (2013). A Comparison Of Partial Least Square Structural Equation Modeling (PLS-SEM) and Covariance Based Structural EquationModeling (CB-SEM) for Confirmatory Factor Analysis International Journal of Engineering Science and Innovative Technology (IJESIT), 2(5), 9.

References [edit]

1. ^ Tenenhaus, M.; Esposito Vinzi, V.; Chatelinc, Y-M.; Lauro, C. (January 2005). "PLS path modeling" (PDF). *Computational Statistics & Data Analysis* **48** (1): 159–205. doi:10.1016/j.csda.2004.03.005.
2. ^ Vinzi, V.; Chin, W.W.; Henseler, J. et al., eds. (2010). *Handbook of Partial Least Squares*. ISBN 978-3-540-32825-4.
3. ^ Tenenhaus, M. (2008). "Component-based structural equation modelling" (PDF).
4. ^ Wold, S; Sjöström, M.; Eriksson, L. (2001). "PLS-regression: a basic tool of chemometrics". *Chemometrics and Intelligent Laboratory Systems* **58** (2): 109–130. doi:10.1016/S0169-7439(01)00155-1.

5. ^ Lindgren, F; Geladi, P; Wold, S (1993). "The kernel algorithm for PLS". *J. Chemometrics* **7**: 45–59. doi:10.1002/cem.1180070104.
6. ^ de Jong, S.; ter Braak, C.J.F. (1994). "Comments on the PLS kernel algorithm". *J. Chemometrics* **8** (2): 169–174. doi:10.1002/cem.1180080208.
7. ^ Dayal, B.S.; MacGregor, J.F. (1997). "Improved PLS algorithms". *J. Chemometrics* **11** (1): 73–85. doi:10.1002/(SICI)1099-128X(199701)11:1<73::AID-CEM435>3.0.CO;2-#.
8. ^ de Jong, S. (1993). "SIMPLS: an alternative approach to partial least squares regression". *Chemometrics and Intelligent Laboratory Systems* **18** (3): 251–263. doi:10.1016/0169-7439(93)85002-X.
9. ^ Rannar, S.; Lindgren, F.; Geladi, P.; Wold, S. (1994). "A PLS Kernel Algorithm for Data Sets with Many Variables and Fewer Objects. Part 1: Theory and Algorithm". *J. Chemometrics* **8** (2): 111–125. doi:10.1002/cem.1180080204.
10. ^ Abdi, H. (2010). "Partial least squares regression and projection on latent structure regression (PLS-Regression)". *Wiley Interdisciplinary Reviews: Computational Statistics* **2**: 97–106. doi:10.1002/wics.51.
11. ^ Trygg, J; Wold, S (2002). "Orthogonal Projections to Latent Structures". *Journal of Chemometrics* **16** (3): 119–128. doi:10.1002/cem.695.
12. ^ Sæbøa, S.; Almøya, T.; Flatbergb, A.; Aastveita, A.H.; Martens, H. (2008). "LPLS-regression: a method for prediction and classification under the influence of background information on predictor variables". *Chemometrics and Intelligent Laboratory Systems* **91** (2): 121–132. doi:10.1016/j.chemolab.2007.10.006.
13. ^ "Package 'pls'" (PDF).

External links [edit]

- imDEV free Excel add-in for PLS and PLS-DA
- PLS in Brain Imaging
- on-line PLS regression (PLSR) at Virtual Computational Chemistry Laboratory
- Uncertainty estimation for PLS
- A short introduction to PLS regression and its history

Authority control GND: 4591652-4

Categories: Regression analysis | Latent variable models | Least squares