Using Partial Least Squares Regression to Analyze Cellular Response Data

Pamela K. Kreeger[1,*]

[1]Department of Biomedical Engineering, University of Wisconsin-Madison Madison, WI 53706

Abstract

This Teaching Resource provides lecture notes, slides, and a problem set for a lecture introducing the mathematical concepts and interpretation of partial least squares regression (PLSR) and were part of a course entitled "Systems Biology: Mammalian Signaling Networks." PLSR is a multivariate regression technique commonly applied to analyze relationships between signaling or transcriptional data and cellular behavior.

Keywords

PLSR; multivariate analysis; systems biology; computational biology; cancer therapy

Introduction

This Teaching Resource is intended for instructors who have familiarity with linear algebra; familiarity with MATLAB will be helpful for the problem assignment. This lecture on partial least squares regression (PLSR) was part of an introductory systems biology course focused on implementation and analysis of systems biology models, which included overviews of several experimental techniques and computational methods. The topic of PLSR followed earlier lectures on the quantitative experimental methods frequently used to gather these data sets (1, 2) and principal component analysis (PCA) (3). PLSR is a multivariate regression technique developed to analyze relationships between independent and dependent variables in large data sets and can therefore be applied to analyze proteomic,

*Corresponding author: Pamela K. Kreeger, Department of Biomedical Engineering, 1550 Engineering Drive, ECB 2154, Madison, WI 53706, Tel. 608-890-2915, Fax 608-265-9239, kreeger@wisc.edu.

transcriptomic, metabolomic, and other cellular data. In particular, PLSR has been used in the systems biology community to analyze relationships between intracellular signals and cellular responses (4-6).

Lecture Notes

Cue-Signal-Response Relationships

Typically, systems biology studies are concerned with analyzing the 'cue-signal-response' relationship, in which cells take in information from their environment (for example, growth factors, extracellular matrix signals, or mechanical stimuli) and this information is propagated through the cellular signaling network to ultimately make cell phenotype decisions (for example, proliferation, death, or differentiation). The cue-signal relationship has frequently been analyzed using mechanistic models, such as mass-action kinetics (7). In contrast, the signal-response relationship is strongly impacted by interactions among multiple pathways and occurs on a longer time scale, making it generally too complex for these detailed mechanistic model forms (Slide 2). For example, fitting of a mass-action kinetic model of the EGFR family that incorporated several receptor and ligand forms, receptor trafficking, and activation of ERK and AKT, resulted in approximately 100 different parameter sets that could be fit equally well to the training data set (8). As more pathways or timescales are added the problem of accurately describing the network topology and estimating parameters becomes even more challenging, making mass-action kinetics impractical. Here, we first briefly discuss univariate approaches to analyze the signal-response relationship, that is how the amount of one signal is used to predict the cellular response (9, 10), and then provide a detailed description of multivariate approaches, such as PLSR, to determine how multiple signaling cascades are integrated in the cellular decision-making process.

Methods to Analyze Signal-Response Relationships

There are several methods to analyze univariate signal-response relationships. This relationship can be described by correlations, such as a 'low' level of signal indicating a certain response and a 'high' level of signal indicating another response (Slide 3). Challenges with this approach include the difficulty in determining cut-offs for the different levels of signal and response. Alternatively, the relationship can be defined more quantitatively using regression analysis. In the example (Slide 3), linear regression of a theoretical signaling value is performed. With the resulting equation, it would be possible to predict the response (y) if given the signaling value (x) for a new condition and conversely, to determine the level of signaling that would be associated with a distinct response.

Although these univariate relationships have been successfully used in various biological systems, it is important to note that they are often insufficient to explain many cellular responses. For example, if the extent of activation of a phosphoprotein is measured under a small range of conditions, it may be possible to conclude that this protein does a particular function. However, as more conditions are measured, this univariate relationship will fail as no single pathway governs cellular decisions alone; indeed, this observation is one of the very cornerstones of the field of systems biology. For example, Janes, *et al.* showed that the

correlation between the amount of phosphorylated c-Jun N-terminal kinase (p-JNK) and apoptosis (a form of cell death) was highly context-dependent (Slide 4): At low doses of epidermal growth factor (EGF) apoptosis increased as p-JNK increased, at high doses of tumor necrosis factor α (TNF-α) apoptosis decreased as p-JNK increased, and at high doses of EGF there was a biphasic relationship (4). One possible interpretation of these results is that p-JNK is not an appropriate signal on which to base this relationship, perhaps because it is not relevant to the downstream response. However, previous studies have suggested a role for JNK in apoptosis. An alternative explanation is that the impact of p-JNK on the apoptotic decision is modified by activity in other pathways and a multivariate model is necessary to accurately interpret this effect. As a result of experimental techniques that allow multiple analytes to be measured simultaneously (known as multiplexing (1)), it is now possible to address the question of how different pathways influence each other; however, the size of the resulting data sets requires mathematical analysis to decode these relationships.

By measuring more signals, responses, or both, a multivariate regression may be performed to develop the signal-response relationship (Slide 5). As an example, multi-linear regression was successfully used to identify the network links between several different stimuli and downstream signaling proteins (11). Although conceptually simple, application of this method is limited, because the number of solutions available depends on the relationship between the number of variables (m) and the number of observations (n). In systems biology, experimental methods that have high degrees of multiplexing are often used. Consequently, there are usually more variables measured than observations, particularly where studies are carried out using different cell lines or patient samples. In this case, there is no unique solution unless the dimensionality of the problem is reduced (Slide 5).

Dimensionality reduction can be accomplished through principal components regression (PCR) or PLSR. In PCR, the matrix of signaling measurements, X, is transformed into principal component space (Slide 6-7). Principal components are a set of orthogonal coordinates that capture the variation in the data matrix and are identified by finding the eigenvectors and associated eigenvalues for the covariance matrix of the data set (for more detail on this method, see (3)). The first component captures the greatest variation, the second component captures the largest fraction of the remaining variation, and so on until the next component does not significantly improve R^2X (the coefficient of determination for that matrix). However, finding the eigenvectors and eigenvalues is computationally intensive for large data matrices. As an alternative, principal components can be derived by the NIPALs (non-linear iterative partial least squares) algorithm, with the original data matrices deconstructed into a sum of vector products for each principal component (12). The two vector products are the loadings vector (p, the directionality of the component, determined by the variables) and the scores vector (t, the magnitude of each observation along that component). In NIPALs, the scores and loadings for the first principal component are determined and then the vector product is subtracted from the initial matrix to find the residual, or unfit data. The next component is then found by analyzing the residual in the same manner (Slide 8). In PCR, the resulting scores matrix is then used to solve the regression problem. However, because the principal components are formulated from only the independent variables, PCR may result in a model that does not accurately describe the signal-response relationship. For example, the largest signaling variation will be captured in

the first principal component but this variation may not correlate with the response measured in Y (Slide 9). For this reason, PLSR is more commonly used to analyze the signal-response relationship, because the principal components are derived based on the covariation between X and Y, rather than the variation in the individual matrices alone.

PLSR

In PLSR, the challenge is to generate the principal components in a manner that emphasizes the variation in X that correlates to the variation in Y (Slide 10). The focus here is on a modification to the NIPALs algorithm that is useful to develop principal components in PLSR (12, 13) (Slide 11). Both the X and Y matrix have associated loadings vectors (p or q) and scores vectors (t or u, the magnitude of each observation along that component). The loadings are determined by the independent and dependent variables, and correspond to the direction of the component in space. A larger loading for variable 1 versus 2 indicates that 1 contributes more strongly to that principal component than 2. The scores vector relates how far each observation projects along that principal component. When NIPALs is used for PLSR, the scores for Y are used to find the loadings for X; In this way, the relationship between X and Y is emphasized for the observations for each component (Slide 12). It is important to note that even though the NIPALs algorithm is similar for PCA and PLSR, the actual components differ due to this exchange of scores step (Slide 13). PCA of a signaling data set will assign the highest loading values to variables that have the highest variation, while in PLSR, the emphasis on co-variation can change this order. For example, when Janes, *et al.* analyzed signaling time courses for AKT, JNK, and MK2 following treatment with various growth factors and cytokines, the variation in AKT was lower than that of JNK and MK2, resulting in a lower loading in principal component 1 of the PCA analysis (14). However, when this same data was analyzed by PLSR to predict the activation of an AKT-substrate, the loading for AKT in principal component 1 increased due to the co-variation between AKT and its downstream target.

Because components are defined sequentially, it is important to determine how many components are needed for the optimal model. Increasing the number of components may improve the model's predictive capability, but this is associated with the cost that the model becomes more challenging to interpret. Additionally, later components are predominated by noise in the experimental data and may actually decrease predictive ability. For this reason, we can analyze three metrics to determine the optimum number of components (Slide 14). R^2X and R^2Y are the coefficients of determination for the X and Y matrices, and Q^2Y is a measure of the predictive ability of the model based on cross-validation (15). The cumulative value for each metric is evaluated after a component is added to the model (Slide 15). By comparing the values for R^2X, R^2Y, and Q^2Y to their values for the previous component, we can determine how much including an additional component improved the fit or predictive ability or both. R^2X and R^2Y will increase with each added component, although the increase becomes much smaller at later components. Because PLSR is primarily used to relate the effects of changes in X on Y, the decision to add a new component is prioritized towards the impact on Q^2Y. If Q^2Y increases significantly the component is retained. The algorithm will then continue until Q^2Y either does not increase significantly or actually decreases.

An important advantage of PLSR is that the developed models can be used to predict the response of data sets that were not included in the training set (Slide 16). Conceptually, this is similar to taking a new value of x and using it in the equation for a line that was fit from a previous data set (Slide 3). However, in PLSR, the regression equation has been found in a new set of dimensions, so the X values for the prediction data set are first transformed into principal-component space (13). The resulting prediction for the response is then transformed back into the original units to determine Y. To evaluate the accuracy of this prediction, DModY (the distance between the model prediction and observation for Y) is calculated (Slide 16). Through this approach it is possible to test in silico the effects of changes in signaling on the resulting cellular response and, thus, to select a lead candidate from a panel of proposed experimental perturbations or to determine how the different X variables impact cellular behavior (16).

An Example of PLSR from the Literature

To demonstrate how PLSR can be used to analyze a systems-level problem, research utilizing PLSR to identify a predictive relationship between the abundance of members of the epidermal growth factor receptor family ErbB signaling network and sensitivity to an ErbB inhibitor in ovarian cancer cells (OvCa) (17). Although receptors of the ErbB type and their ligands are present in many ovarian cancers, only a subset of patients have responded to ErbB-targeted therapies in clinical trials (18). Therefore, to maximize the benefit of these targeted therapies, methods to effectively predict which patients will respond prior to initiating treatment are needed (Slide 17). Although prior studies have identified univariate biomarkers (for example overexpression of specific genes in the pathway or mutations in those genes) that predict response to these drugs in other tumor types, these markers have not been able to predict therapeutic response in ovarian cancer. Because the ErbB system is composed of four receptors that homo- and hetero-dimerize in response to stimulation by a family of thirteen ligands, Prasasya, *et al.* utilized PLSR to examine if the quantitative balance of the abundance of these ErbB network components could be used to determine which ovarian cancer cells would be sensitive to canertinib (CI-1033), a pan-ErbB inhibitor (17) (Slide 18).

To examine this question, the amounts of the four receptors (ErbB1-4) and three ligands (TGFα, NRG1β, and HB-EGF) were quantified across a panel of six ovarian cancer lines using quantitative Western blots and ELISAs (Slide 19). Sensitivity to canertinib was determined by a cytotoxicity assay. Although the effective concentration for a 50% response (EC50) for this drug varied only slightly, some cell lines showed a much greater proportion of cell death than others (Slide 20). Therefore, the sensitivity value was determined as the difference between the baseline amount of cell death and the amount at the maximum canertinib dose tested. The resulting data matrix was mean-centered and unit-variance scaled prior to analysis by PLSR (1). By examining the values of R^2Y and Q^2Y for the first three components, a two-component model was selected for further evaluation (inclusion of the third component increased both R^2Y and Q^2Y, but did not increase Q^2Y significantly). A plot of model predictions versus experimental observations demonstrated that this model accurately predicted the training data set (Slide 21).

Following the development of the PLSR model, it can be very informative to analyze the values of both the scores and loadings to probe the relationship between the independent and dependent variables. As stated above, the scores relate how far each observation projects along that principal component. In the case of Prasasya, *et al.*, a plot of the scores for each sample for principal component 1 and principal component 2 (Slide 22) can be analyzed in conjunction with the raw data from the sensitivity analysis (Slide 20). From this, we see that the first component captures most of the variation responsible for predicting drug sensitivity: The most sensitive cell line projects the farthest along principal component 1, and the least sensitive cell line projects negatively in principal component 1 (Slide 22). The second component fine-tunes the prediction, particularly for the four cell lines with similar sensitivity. For example, OVCA432 has a more negative projection along principal component 1 than OVCA429, even though OVCA432 was more sensitive; this is corrected by the different projections in principal component 2. This interpretation is further strengthened by examining the loadings plot (Slide 23), where the loading for canertinib sensitivity (CI-1033) is a positive value for both principal components. The values for the loadings for independent and dependent variables can then be examined to determine how these variables contribute to the predicted sensitivity. For example, the loadings plot for this model suggests that the amount of ErbB1 is negatively correlated with sensitivity (negative value in both components), and that NRG1β is positively correlated with sensitivity (positive value in both components, Slide 23). This is consistent with the observation that the cell line with the highest sensitivity (OVCA420) has the highest abundance of NRG1β, and the cell line with the weakest sensitivity (OVCAR5) has the highest abundance of ErbB1. However, these observations also demonstrate the importance of deriving a multivariate relationship. Although OVCA429 had only slightly less ErbB1 than did OVCAR5 but OVCA429 was nearly 35% more sensitive (Slide 19, 20).

To further analyze the composition of this model, the VIPs (Variable Importance of Projections) were examined for each of the X variables. In contrast to loadings and scores, VIPs are evaluated across the entire model. The VIP is the sum over all of the components of the contribution of that variable. These contributions are determined based on the weight for the variable in each component in combination with the variation captured in that component versus the other components. The average of the VIPs for all variables is 1 in PLSR; therefore, values greater than 1 indicate variables that are more important for the model. In the model presented in (17), the most important variables were NRG1β and ErbB1 (Slide 24).

The model was next used to predict the response of six additional cell lines that were not included in the training set (Slide 25). To test the predictive capacity of the two-component model, the abundance of receptors and ligands and the sensitivity to canertinib-induced cell death were experimentally determined for six additional ovarian cancer cell lines. The model predicted the sensitivity of three of these cell lines with reasonable accuracy; however, there were large differences between the observed and predicted values for the other three cell lines, resulting in prediction of a negative value for sensitivity for two cell lines. This would be interpreted as a decrease in cell death following treatment with canertinib, an outcome that was not seen in any cell line tested.

To determine if the model could be modified to improve prediction accuracy, the X data was examined for the training and prediction data set. The amount of ErbB2 for the cell lines that were inaccurately predicted was out of the range of the training data set (Slide 26); therefore, Prasasya, *et al.* examined if expanding the training data set to include a cell line with ErbB2 abundance in this range would improve prediction accuracy, as prior studies have demonstrated that the predictive capacity of PLSR models is dependent upon the training data set incorporating the full range of values from the prediction data set (17, 19). This modification improved the prediction accuracy of the model (Slide 27). Additionally, Prasasya, *et al.* examined if excluding ErbB2 from the X block would improve prediction accuracy (17). This modification also improved the model prediction accuracy (Slide 28). The observation that leaving out data could improve model accuracy was then further examined by testing the 127 possible combinations of X matrices (that is X that included combinations of 1 to 7 variables) (Slide 29). From this analysis, the nine best models were observed to include ErbB1 and NRG1β, consistent with the high VIPs for these variables in the full 7-variable model. Additionally, none of the high-ranking models included ErbB2, consistent with the problems observed when including ErbB2 in the full 7-variable model. Interestingly, the best models all included two or three of the ErbB ligands. Increased abundance of ligands would be expected to increase ErbB pathway activity, which could make cells more sensitive to ErbB-targeted therapy. The inclusion of ligand measurements in a model to predict therapeutic response contrasts with current clinical tests that primarily measure receptor abundance. Although this paper demonstrates the interpretation of many of the metrics used in PLSR analysis, other papers could be substituted or added to emphasize other elements of PLSR. For example, Janes, *et al.* demonstrated the utility of PLSR to develop hypotheses about biological mechanisms (4).

Concluding Thoughts

Other versions of PLSR have been developed that may be appropriate for analysis of large data sets (Slide 30). For example, discriminant PLS (DPLS) is a variant that can be used when the values in the response matrix are classifications (20). Orthogonal PLS (OPLS) models the variation in X into two parts – one that is related to Y and one that is not related to Y (21). Although OPLS does not change predictive power, it may simplify model interpretation. Finally, O2PLS splits the variation into 3 parts: The variation in X that is unrelated to Y, the variation in Y that is unrelated to X, and the variation in X that is related to Y. Complete details on these methods can be presented in subsequent lectures.

Problem Set

Information for the Instructor

PLSR is available through various commercial packages - for example, SIMCA-P (Umetrics), Unscrambler X (CAMO), and XLSTAT (Addinsoft); the problem set solution utilizes a publically available MATLAB script (http://www.cdpcenter.org/files/plsr/nipals.html – this website also has a free example problem using this script which may help students to complete this problem). Small deviations in results are possible between different platforms due to variations in the algorithm used. Depending on their expertise with

MATLAB, students may find some of these steps easier to perform in Excel or other software.

Questions for the Students

For a set of four samples you measure the abundance of 4 phosphorylated proteins at 10 different times and the proportion of cell death for these conditions at 3 different times. Analyze this data compendium (Samples 1-4 in the supplementary materials PLSdata.xls) by PLSR. In the data file, each row indicates a unique sample, while the values in the columns represent the values for the associated signaling time points (X matrix) and percentage of cells that were dead (Y matrix).

1. Unit variance scale the raw data for both the X and Y matrices.

2. Use the scaled X and Y matrices to construct a 1-component PLSR model and a 2-component PLSR model.

3. Plot the observed versus predicted values for both PLSR models, converting the scaled predictions back to original units prior to plotting.

4. Report the R^2Y value for each model.

5. For the 2-component model, plot the scores for each observation (PLSdata.xls). How would you interpret these components given the information about the samples provided in Table 1?

6. For the 2-component model, plot the loadings for both the X and Y variables.

7. You test a new sample (Sample 5 in PLSdata.xls). Using the 2-component model, predict what the response will be and compare it to the measured values.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

1. Albeck JG, MacBeath G, White FM, Sorger PK, Lauffenburger DA, Gaudet S. Collecting and organizing systematic sets of protein data. Nat Rev Mol Cell Biol. 2006; 7:803–812. [PubMed: 17057751]

2. Schulze A, Downward J. Navigating gene expression using microarrays--a technology review. Nat Cell Biol. 2001; 3:E190–195. [PubMed: 11483980]

3. Clark NR, Ma'ayan A. Introduction to statistical methods to analyze large data sets: Principal components analysis. Sci Signal. 2011; 4:tr3. [PubMed: 21917717]

4. Janes KA, Albeck JG, Gaudet S, Sorger PK, Lauffenburger DA, Yaffe MB. A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. Science. 2005; 310:1646–1653. [PubMed: 16339439]

5. Kumar D, Srikanth R, Ahlfors H, Lahesmaa R, Rao KV. Capturing cell-fate decisions from the molecular signatures of a receptor-dependent signaling response. Mol Syst Biol. 2007; 3:150. [PubMed: 18059445]

6. Kreeger PK, Mandhana R, Alford SK, Haigis KM, Lauffenburger DA. Ras mutations affect tumor necrosis factor-induced apoptosis in colon carcinoma cells via erk-modulatory negative and positive feedback circuits along with non-erk pathway effects. Cancer Res. 2009; 69:8191–8199. [PubMed: 19789336]

7. Aldridge BB, Burke JM, Lauffenburger DA, Sorger PK. Physicochemical modelling of cell signalling pathways. Nat Cell Biol. 2006; 8:1195–1203. [PubMed: 17060902]

8. Chen WW, Schoeberl B, Jasper PJ, Niepel M, Nielsen UB, Lauffenburger DA, Sorger PK. Input-output behavior of erbb signaling pathways as revealed by a mass action model trained against dynamic data. Mol Syst Biol. 2009; 5:239. [PubMed: 19156131]

9. Schoeberl B, Pace EA, Fitzgerald JB, Harms BD, Xu L, Nie L, Linggi B, Kalra A, Paragas V, Bukhalid R, Grantcharova V, Kohli N, West KA, Leszczyniecka M, Feldhaus MJ, Kudla AJ, Nielsen UB. Therapeutically targeting erbb3: A key node in ligand-induced activation of the erbb receptor-pi3k axis. Sci Signal. 2009; 2:ra31. [PubMed: 19567914]

10. Wang Z, Birch CM, Sagotsky J, Deisboeck TS. Cross-scale, cross-pathway evaluation using an agent-based non-small cell lung cancer model. Bioinformatics. 2009

11. Alexopoulos LG, Saez Rodriguez J, Cosgrove BD, Lauffenburger DA, Sorger PK. Networks inferred from biochemical data reveal profound differences in tlr and inflammatory signaling between normal and transformed hepatocytes. Mol Cell Proteomics. 2010

12. Geladi P, Kowalski BR. Partial least-squares regression: A tutorial. Anal Chim Acta. 1986; 185:1–17.

13. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Trygg, J.; Wikstrom, C.; Wold, S. Multi- and megavariate data analysis part i: Basic principles and applications. Umetrics Academy; 2006. p. 425

14. Janes KA, Yaffe MB. Data-driven modelling of signal-transduction networks. Nat Rev Mol Cell Biol. 2006; 7:820–828. [PubMed: 17057752]

15. Efron, B.; Tibshirani, R. An introduction to the bootstrap. CRC Press; Boca Raton: 1993.

16. Kreeger PK, Wang Y, Haigis KM, Lauffenburger D. Integration of multiple signaling pathway activities resolves k-ras / n-ras mutation paradox in colon epithelial cell response to inflammatory cytokine stimulation. Integrative Biology. 2010; 2:202–208. [PubMed: 20473400]

17. Prasasya RD, Vang KZ, Kreeger PK. A multivariate model of erbb network composition predicts ovarian cancer cell response to canertinib. Biotechnol Bioeng. 2012; 109:213–224. [PubMed: 21830205]

18. Yap TA, Carden CP, Kaye SB. Beyond chemotherapy: Targeted therapies in ovarian cancer. Nat Rev Cancer. 2009; 9:167–181. [PubMed: 19238149]

19. Miller-Jensen K, Janes KA, Brugge JS, Lauffenburger DA. Common effector processing mediates cell-specific responses to stimuli. Nature. 2007; 448:604–608. [PubMed: 17637676]

20. Lau KS, Juchheim AM, Cavaliere KR, Philips SR, Lauffenburger DA, Haigis KM. In vivo systems analysis identifies spatial and temporal aspects of the modulation of tnf-alpha-induced apoptosis and proliferation by mapks. Sci Signal. 2011; 4:ra16. [PubMed: 21427409]

21. Cosgrove BD, Alexopoulos LG, Saez-Rodriguez J, Griffith LG, Lauffenburger DA. A multipathway phosphoproteomic signaling network model of idiosyncratic drug- and inflammatory cytokine-induced toxicity in human hepatocytes. Conf Proc IEEE Eng Med Biol Soc. 2009; 2009:5452–5455. [PubMed: 19964679]

Table 1

Observations

	Cell Type	Treatment
Sample 1	A	Vehicle
Sample 2	A	Drug
Sample 3	B	Vehicle
Sample 4	B	Drug