Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression

Saikat Maitra and Jun Yan

Abstract: Dimension reduction is one of the major tasks for multivariate analysis, it is especially critical for multivariate regressions in many P&C insurance-related applications. In this paper, we'll present two methodologies, principle component analysis (PCA) and partial least squares (PLC), for dimension reduction in a case that the independent variables used in a regression are highly correlated. PCA, as a dimension reduction methodology, is applied without the consideration of the correlation between the dependent variable and the independent variables, while PLS is applied based on the correlation. Therefore, we call PCA as an unsupervised dimension reduction methodology, and call PLS as a supervised dimension reduction methodology. We'll describe the algorithms of PCA and PLS, and compare their performances in multivariate regressions using simulated data.

Key Words: PCA, PLS, SAS, GLM, Regression, Variance-Covariance Matrix, Jordan Decomposition, Eigen Value, Eigen Factors.

Introduction

In large-scale data mining and predictive modeling, especially for multivariate regression exercises, we often start with a large number of possible explanatory/predictive variables. Therefore, variable selection and dimension reduction is a major task for multivariate statistical analysis, especially for multivariate regressions. A well-known method in regression analysis for dimension reduction is called stepwise regression algorithm, which is covered by many statistical softwares such as SAS and SPSS. One of the major limitations of the algorithm is that when several of the predictive variables are highly correlated, the tests of statistical significance that the stepwise method is based on are not sound, as independence is one of the primary assumptions of these tests.

Often, many variables used as independent variables in a regression display a high degree of correlation, because those variables might be measuring the same characteristics. For example, demographic variables measuring population density characteristics or weather characteristics are often highly correlated.

A high degree of correlation among the predictive variables increases the variance in estimates of the regression parameters. This problem is known as multi-colinearity in regression literature (Kleinbaum et al. [4]). The parameter estimates in a regression equation may change with a slight change in data and hence are not stable for predicting the future.

In this paper, we will describe two methodologies, principle component analysis (PCA) and partial least square (PLS), for dimension reduction in regression analysis when some of the independent variables are correlated. We'll describe what algorithm is used in each methodology and what the major differences are between the two methodologies.

Principal Component Analysis

PCA is a traditional multivariate statistical method commonly used to reduce the number of predictive variables and solve the multi-colinearity problem (Bair et al. [3]). Principal component analysis looks for a *few* linear combinations of the variables that can be used to summarize the data without losing too much information in the process. This method of dimension reduction is also known as "parsimonious summarization" (Rosipal and Krämer [6]) of the data. We will now formally define and describe principal components and how they can be derived. In the process we introduce a few terms for the sake of completeness.

Data Matrix

Let $Xn\times p$ denote the matrix of predictive variables (henceforth referred to as *data-matrix*), where each row denotes an observation on p different predictive variables, $X1,X2, \dots , Xp$. We will denote a random observation from this matrix by $x1\times p$. The problem at hand is to select a subset of the above columns that holds most of the information.

Variance-Covariance Matrix

Let σ_{ij} denote the co-variance between X_i and X_j in the above data-matrix. We will denote the matrix of $((\sigma_{ij}))$ by $\sum$. Note that the diagonal elements of $\sum$ are the variances of X_i. In actual calculations σ_{ij}s may be estimated by their sample counterpart's s_{ij} or sample covariance calculated from the data. The matrix of standard deviations $((s_{ij}))$ will be denoted by S. Note both $\sum$ and S are $p\times p$ square and symmetric matrices.

Linear Combination

A linear combination of a set of vectors $(X1,X2, \dots , Xp)$ is an expression of the type $\sum\alpha iXi$ (i=1 to p) and αis are scalars. A linear combination is said to be normalized or standardized if $\sum|\alpha i|=1$ (sum of absolute values). In the rest of the article, we will refer to the standardized linear combination as SLC.

Linear Independence

A set of vectors are said to be linearly independent if none of them can be written as a linear combination of any other vectors in the set. In other words, a set of vectors $(X_1, X_2, \dots , X_p)$ is linearly independent if the expression $\sum \alpha_i X_i = 0 \rightarrow \alpha_i = 0$ for all values of *i*. A set of vectors not linearly independent is said to be linearly dependent.

Statistically, correlation is a measure of linear dependence among variables and presence of highly correlated variables indicate a linear dependence among the variables.

Rank of a Matrix

Rank of a matrix denotes the maximum number of linearly independent rows or columns of a matrix. As our data-matrix will contain many correlated variables that we seek to reduce, rank of the data-matrix, $X_{n\times p}$, is less than or equal to p.

Jordan Decomposition of a Matrix

Jordan decomposition or spectral decomposition of a symmetric matrix is formally defined as follows.

Any symmetric matrix $A_{p\times p}$ can be written as $A = \Gamma \Lambda \Gamma^{T} = \sum \lambda_i \gamma'_{(i)} \gamma_{(i)}$ where $\Lambda_{p\times p}$ is a diagonal matrix with all elements 0 except the diagonal elements and $\Gamma_{p\times p}$ is an orthonormal matrix, i.e., $\Gamma\ \Gamma' = I$ (identity matrix).

The diagonal elements of Λ are denoted by λ_i (i=1 to p) and the columns of Γ are denoted by $\gamma_{(i)}$ (i=1 to p). In matrix algebra, λ_is are called eigen values of A and $\gamma_{(i)}$s are the corresponding eigen vectors.

If A is not a full rank matrix, i.e., rank(A) = $r < p$, then there are only r non-zero eigen values in the above Jordan decomposition, with the rest of the eigen values being equal to 0.

Principal Components

In principal component analysis, we try to arrive at a suitable SLC of the data-matrix X based on the Jordan decomposition of the variance-covariance matrix $\sum$ of X or equivalently based on the correlation matrix Φ of X. We denote the mean of the observations as $\mu_{1\times p}$.

Let $x_{1\times p} = (x_1, x_2, \dots, x_p)$ denote a random vector observation in the data-matrix (i.e., transpose of any row of the $n\times p$ data matrix), with mean $\mu_{1\times p}$ and covariance matrix $\sum$. A principal component is a transformation of the form $x_{1\times p} \rightarrow y_{1\times p} = (x\text{-}\mu)_{1\times p}\ \Gamma_{p\times p}$, where Γ is obtained from the Jordan decomposition of $\sum$, i.e., $\Gamma^{T} \sum \Gamma = \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ with λ_is being the eigen values of the decomposition. Each element of $y_{1\times p}$ is a linear combination of the elements of $x_{1\times p}$. Also each element of y is independent of the other.

Thus we obtain p independent principal components corresponding to the p eigen values of the Jordan decomposition of $\sum$. Generally, we will only use the first few of these principal components for a regression. In the next section, we will list the major properties of the principal components as obtained above. This will help us to understand why the first few of the principal components may hold the majority of the information and thus help us reduce the dimension in a regression without losing too much information.

Properties of Principal Components (Anderson [2])

The following result justifies the use of PCA as a valid variable reduction technique in regression problems, where a first few of the principal components are used as predictive variables.

Let x be a random p dimensional vector with mean μ and covariance matrix $\sum$. Let y be the vector of principal components as defined above. Then the following holds true.

(i) $E(y_i) = 0$

(ii) $\mathrm{Var}(y_i) = \lambda_i$

(iii) $\mathrm{Cov}(y_i,y_j) = 0$

(iv) $\mathrm{Var}(y_1) \geq \mathrm{Var}(y_2) \geq \ldots \geq \mathrm{Var}(y_p)$

(v) No SLC of x has variance larger than λ_1, the variance of the first principal component.

(vi) If $z=\sum\alpha_i x_i$ be a SLC of x, which is uncorrelated with first k principal components, then variance of z is maximized if z equals the $(k+1)^{\text{th}}$ principal component.

Item (iii) above justifies why using principal components instead of the raw predictive variables will remove the problem of multi-colinearity.

Items (iv), (v), and (vi) indicate that principal components successively capture the maximum of the variance of x and that there is no SLC that can capture maximum variance without being one of the principal components. When there is high degree of correlation among the original predictive variables, only the first few of the principal components are likely to capture majority of the variance of the original predictive variables. The magnitude of λ_is provides the measure of variance captured by the principal components and should be used to select the first few components for a regression.

A Numerical Example of PCA

In this section we describe the process of building principle components in a multivariate regression set up using a simulated data for line of business of business owners policies (BOP). The simulated data has been used for the 2006 and 2007 CAS Limited Attendance Predictive Modeling Seminars.

Description of the Data

The simulated data is in a policy-year level. That means each data record contains information of a twelve-month BOP policy. In the data, we simulated claim frequency, claim count over per $000 premium, and six correlated policy variables.

The policy variables in this example are:

fireProt – Fire Protection Class

numBldg – Number of Building in Policy

numLoc – Number of Locations in Policy

bldgAge – Maximum Building Age

bldgContents – Building Coverage Indicator

polage – Policy Age

All the predictive variables are treated as continuous variables including the bldgContents variable. Both the multivariate techniques described in this paper works only with continuous and ordinal variables. Categorical variables cannot be directly analyzed by these methods for variable reduction.

The correlation matrix of the above predictive variables is:

	fireProt	**numBldg**	**numLoc**	**bldgAge**	**bldgContents**	**polAge**
fireProt	1.0000	-0.3466	0.0020	0.2921	-0.0945	-0.0328
numBldg	-0.3466	1.0000	0.8012	-0.2575	0.1216	0.0494
numLoc	0.0020	0.8012	1.0000	-0.0650	0.0619	0.0417
bldgAge	0.2921	-0.2575	-0.0650	1.0000	-0.0694	0.0287
bldgContents	-0.0945	0.1216	0.0619	-0.0694	1.0000	0.0068
polAge	-0.0328	0.0494	0.0417	0.0287	0.0068	1.0000

As can be seen, numBldg and numLoc are highly correlated and the variable fireProt has significant correlation with two other variables.

Principle Components Regression

As described earlier, the principle components are obtained by eigen-value decomposition of the covariance or correlation matrix of the predictive variables under consideration. Generally, most statistical software can compute the principle components once we specify the data set and the variables with which we want to construct principle components. SAS, for example, provides outputs of the linear coefficients (eigen vectors) along with mean and standard deviations of each predictive variables. These can be used to compute the principle components on the data set for regression.

Step 1: Compute the Jordan decomposition of the correlation matrix and obtain the eigen vector ($\Gamma i=\{\gamma 1i, \gamma 2i,\ldots, \gamma pi\}$) corresponding to each eigen value (λi).

The six eigen values of the eigen value decomposition of the above correlation matrix are as follows:

Eigen Values	**Proportion of Total**	**Cumulative Proportion of Total**
2.00767648	0.294943617	0.294943617
1.9965489	0.293308887	0.588252504
1.00066164	0.147005141	0.735257644
0.96103098	0.141183082	0.876440726
0.71945588	0.105693782	0.982134508
0.12161012	0.017865492	1

As we can see the first four eigen values capture about 90% of the information in the correlation matrix.

The eigen vectors (columns of matrix Γ in the Jordan decomposition) corresponding to each of the eigen values above are:

Eigen Vector 1	**Eigen Vector 2**	**Eigen Vector 3**	**Eigen Vector 4**	**Eigen Vector 5**	**Eigen Vector 6**
(0.336140)	0.589132	(0.135842)	0.167035	0.654102	0.256380
0.664985	0.178115	(0.053062)	(0.050656)	(0.097037)	0.715033
0.561060	0.501913	(0.109841)	0.005781	0.065075	(0.645726)
(0.313430)	0.558248	0.087962	0.212191	(0.729197)	0.075033
0.168213	(0.204757)	0.127973	0.953512	0.061786	0.020003
0.059014	0.125363	0.970851	(0.123655)	0.151504	0.002265

Step 2: Construct the principle components corresponding to each eigen value by linearly combining the standardized predictive variables using the corresponding eigen vector.

Hence the first principle component can be computed as:

PrinComp1 =

-0.336139581 * (fireProt - 4.55789) / 2.4533790858

+ 0.6649848702 * (numBldg - 1.10179) / 0.6234843087

+ 0.5610599572 * (numLoc - 1.16947) / 0.4635645241

+ -0.313430401 * (bldgAge - 48.5329) / 17.719473959

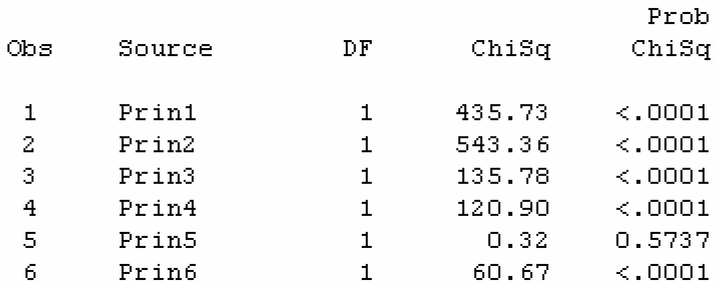
+ 0.1682134808 * (bldgContents - 2.36607) / 0.8750945166

+ 0.0590138772 * (polage - 4.81878) / 3.1602055599

Note that each variable is standardized while computing the principal components.

Now, we'll use the principle components we constructed above in a generalized linear model (GLM) type of regression. There are lot of papers and presentations on GLM ([1], [5]), and we will not spend effort here to describe the related concepts and details. The only two characteristics of GLM that we like to mention are error distribution and link function. Unlike the traditional ordinary regressions, a GLM can select any distribution within the exponential family as the model for the distribution of the target variable. GLM also allows us to use a non-linear link function that permits us to incorporate a non-linear relationship between the target variables and the predictive variables. For example, while fitting a severity curve often the LOG of the loss value can be modeled more easily than the actual loss value in a linear model. GLM allows us to accomplish this by specifying a LOG as the link function. However, it is to be noted that GLM is still linear in terms of the regression parameters.

In this numerical example for PCA, we choose Poisson distribution for regression error and choose IDENTIT*Y* as a link function. We used the claim frequency, claim count over $000 premium, as the dependent variable and used the principle components constructed above as independent variables. The summary of the regression is displayed below:



```
                                                 Prob
Obs     Source          DF         ChiSq        ChiSq

  1     Prin1            1        435.73       <.0001
  2     Prin2            1        543.36       <.0001
  3     Prin3            1        135.78       <.0001
  4     Prin4            1        120.90       <.0001
  5     Prin5            1          0.32       0.5737
  6     Prin6            1         60.67       <.0001
```

The *P*-values and chi-square-statistics demonstrate that the first three principle components explained about 75% the predictive power of the original six policy variables. But, we also noticed the rank of the predictive power didn't line up with the order of the principle components. For example, the first principle component is less explanatory for the target than the second, even though the first principle component contains more information on the six original policy variables. In the next section, we'll describe another dimension reduction technique, partial least squares (PLS), which can be used to solve the problem.

PARTIAL LEAST SQUARES

In the last section we discussed applying PCA in regression as a dimension reduction technique as well as using it to deal with multi-colinearity problems. One drawback of PCA technique in its original form is that it arrives at SLCs that capture only the characteristics of the *X*-vector or predictive variables. No importance is given to how each predictive variable may be related to the dependent or the target variable. In a way it is an unsupervised dimension reduction technique. When our key area of application is multivariate regression, there may be considerable improvement if we build SLCs of predictive variables to capture as much information in the raw predictive variables as well as in the relation between the predictive and target variables. Partial least square (PLS) allows us to achieve this balance and provide an alternate approach to PCA technique. Partial least squares have been very popular in areas like chemical engineering, where predictive variables often consist of many different measurements in an experiment and the relationships between these variables are ill-understood (Kleinbaum et al. [4]). These measurements often are related to a few underlying latent factors that remain unobserved. In this section, we will describe PLS technique and discuss how it can be applied in regression problems by demonstrating it on our sample data.

Description of the Technique

Assume X is a $n\times p$ matrix and Y is a $n\times q$ matrix. The PLS technique works by successively extracting factors from both X and Y such that covariance between the extracted factors is maximized. PLS method can work with multivariate response variables (i.e., when Y is an $n\times q$ vector with q>1). However, for our purpose we will assume thatwe have a single response (target) variable i.e., Y is $n\times 1$ and X is $n\times p$, as before.

PLS technique tries to find a linear decomposition of X and Y such that $X = TP^T + E$ and $Y=UQ^T + F$, where

$T\ n\times r = X$-scores $U\ n\times r = Y$-scores

P $p\times r$= X-loadings Q $1\times r$ = Y-loadings

E $n\times p$ = X-residuals F $n\times 1$ = Y-residuals (1)

Decomposition is finalized so as to maximize covariance between T and U. There are multiple algorithms available to solve the PLS problem. However, all algorithms follow an iterative process to extract the X-scores and Y-scores.

The factors or scores for X and Y are extracted successively and the number of factors extracted (r) depends on the rank of X and Y. In our case, Y is a vector and all possible X factors will be extracted.

Eigen Value Decomposition Algorithm

Each extracted x-score are linear combinations of X. For example, the first extracted x-score t of X is of the form t=Xw, where w is the eigen vector corresponding to the first eigen value of X^TYY^TX. Similarly the first y-score is u=Yc, where c is the eigen vector corresponding to the first eigen value of Y^TXX^TY. Note that X^TY denotes the covariance of X and Y.

Once the first factors have been extracted we deflate the original values of X and Y as,

X_1=X – tt^TX and Y_1=Y- tt^TY. (2)

The above process is now repeated to extract the second PLS factors.

The process continues until we have extracted all possible latent factors t and u, i.e., when X is reduced to a null matrix. The number of latent factors extracted depends on the rank of X.

A NUMERICAL EXAMPLE FOR PLS

In this section we will illustrate how to use the PLS technique to obtain X-scores that will then be used in regression. The data we used for this numerical example is the same as we used for the last numerical example of PCA. The target variable and all the predictive variables used in the last numerical example will be also used in this numerical example.

Partial Least Squares

As we described in the last section, PLS tries to find a linear decomposition of X and Y such that $X=TP^T+E$ and $Y=UQ^T+F$, where

T = X-scores	U = Y-scores
P = X-loadings	Q = Y-loadings
E = X-residuals	F = Y-residuals

Decomposition is finalized so as to maximize covariance between T and U. The PLS algorithm works in the same fashion whether Y is single response or multi-response.

Note that the PLS algorithm automatically predicts Y using the extracted Y-scores (U). However, our aim here is just to obtain the X-scores (T) from the PLS decomposition and use them separately for a regression to predict Y. This provides us the flexibility to use PLS to extract orthogonal factors from X while not restricting ourselves to the original model of PLS.

Unlike PCA factors, PLS factors have multiple algorithms available to extract them. These algorithms are all based on iterative calculations. If we use the eigen value decomposition algorithm discussed earlier, the first step is to compute the covariance X^TY. The covariance between the six predictive variables and the target variable are:

2,208.72
9,039.18
9,497.47
2,078.92
2,858.97
(2,001.69)

As noted, the first PLS factor can be computed from the eigen value decomposition of the matrix X^TYY^TX. The X^TYY^TX matrix is:

4,878,441	19,965,005	20,977,251	4,591,748	6,314,657	(4,421,174)
19,965,005	817,067,728	85,849,344	18,791,718	25,842,715	(18,093,644)
20,977,251	85,849,344	90,201,995	19,744,478	27,152,967	(19,011,011)
4,591,748	18,791,718	19,744,478	4,321,904	5,943,562	(4,161,355)
6,314,657	25,842,715	27,152,967	5,943,562	8,173,695	(5,722,771)
(4,421,174)	(18,093,644)	(19,011,011)	(4,161,355)	(5,722,771)	4,006,769

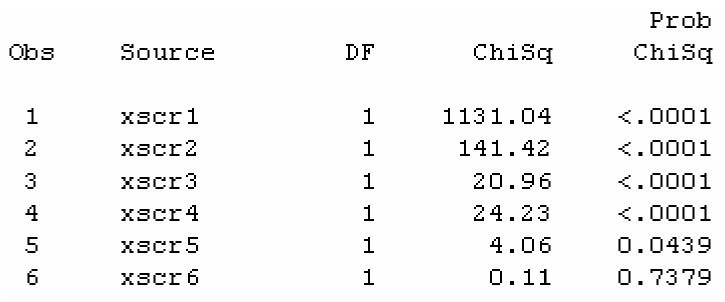
The first eigen vector of the eigen value decomposition of the above matrix is:

{ -0.1588680, -0.6501667, -0.6831309, -0.1495317, -0.2056388, 0.1439770}.

The first PLS *X*-scrore is determined by linearly combining the predictive variables using the above values.

$$\begin{aligned}
Xsr1 = \;& - 0.1588680 * (\text{fireProt} - 4.55789) / 2.4533790858 \\
& - 0.6501667 * (\text{numBldg} - 1.10179) / 0.6234843087 \\
& - 0.6831309 * (\text{numLoc} - 1.16947) / 0.4635645241 \\
& - 0.14953171 * (\text{bldgAge} - 48.5329) / 17.719473959 \\
& - 0.2056388 * (\text{bldgContents} - 2.36607) / 0.8750945166 \\
& + 0.1439770 * (\text{polage} - 4.81878) / 3.1602055599
\end{aligned}$$

Once the first factor has been extracted, the original *X* and *Y* is deflated by an amount ($Xscr1*Xscr^T$) times the original *X* and *Y* values. The eigen value decomposition is then performed on the deflated values, until all factors have been extracted (refer to formula 2).



```
                                                Prob
Obs     Source         DF       ChiSq          ChiSq

 1      xscr1           1     1131.04         <.0001
 2      xscr2           1      141.42         <.0001
 3      xscr3           1       20.96         <.0001
 4      xscr4           1       24.23         <.0001
 5      xscr5           1        4.06         0.0439
 6      xscr6           1        0.11         0.7379
```

We next perform a GLM using the same claim frequency as the dependent variable and the six PLS components, *xscr*1 – *xscr*6, as independent variables. Same as we did in the numerical example for PCA, we still choose Poisson distribution for error the term and an IDENTIT*Y* link function. The regression statistics are displayed below.

Comparing to the ChiSq statistics derived from the GLM using PCA, we can see how each PLS factors are extracted in order of significance and predictive power.

Further Comparison of PCA and PLS

In this section, we have done a simulation study to compare principal components method against the partial least squares methods as a variable reduction technique in regression. A number of simulated datasets were created by re-sampling from original data. PCA and PLS analysis were performed on these data samples and ChiSq statistics of the extracted PCA factors and PLS factors were compared. The exhibit below shows the results on three such samples.

	Simulated Sample 1		**Simulated Sample 2**		**Simulated Sample 3**	
Extracted Factor #	**ChiSq Statistics for PCA Factors**	**ChiSq Statistics for PLS Factors**	**ChiSq Statistics for PCA Factors**	**ChiSq Statistics for PLS Factors**	**ChiSq Statistics for PCA Factors**	**ChiSq Statistics for PLS Factors**
1	79.79	190.73	71.62	160.35	51.44	144.03
2	101.65	24.55	65.18	25.61	43.28	19.21
3	4.78	9.06	34.73	7.72	35.99	0.53
4	17.19	3.58	4.61	5.13	22.65	1.86
5	0.75	0.44	0.21	0.24	2.11	1.16
6	17.91	0.3	20.29	0.14	4.66	0.15

We can see from the above table that the chi-squared statistics of the first two PLS factors are always more than the corresponding two PCA factors in capturing more information.

Summary

PCA and PLS serve two purposes in regression analysis. First, both techniques are used to convert a set of highly correlated variables to a set of independent variables by using linear transformations. Second, both of the techniques are used for variable reductions. When a dependent variable for a regression is specified, the PLS technique is more efficient than the PCA technique for dimension reduction due to the supervised nature of its algorithm.

References

[1] Anderson, D., et al., "A Practitioner's Guide to Generalized Linear Models," CAS Discussion Paper Program (Arlington, Va.: Casualty Actuarial Society, 2004).

[2] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd Edition (New York: John Wiley & Sons, 1984).

[3] Bair, Eric, Trevor Hastie, Paul Debashis, and Robert Tibshirani, "Prediction by Supervised Principal Components," *Journal of the American Statistical Association* 101, no. 473, 2006, pp. 119-137(19).

[4] Kleinbaum, David G., et al., *Applied Regression Analysis and Multivariable Methods,* 3rd Edition (Pacific Grove, Ca.: Brooks/Cole Publishing Company, 1998).

[5] McCullagh, P. and J.A Nelder, *Generalized Linear Models,* 2nd Edition (London: Chapman and Hall, 1989).

[6] Rosipal, Roman, and Nicole Krämer, "Overview and Recent Advances in Partial Least Squares," in *Subspace, Latent Structure and Feature Selection*, Saunders, C., et al. (eds.) (Heidelberg: Springer-Verlag, 2006) vol. 3940, pp. 34-51.