

A Fast Parsimonious Maximum Likelihood Approach for Predicting Outcome Variables from a Large Number of Predictors

Jay Magidson

Statistical Innovations Inc.
Belmont, Massachusetts, United States jay@statisticalinnovations.com

Abstract. A new model with K correlated components is presented for predicting outcome variables where the number of predictors G may exceed the total sample size N . A fast maximum likelihood algorithm provides closed-form expressions for the model parameters and statistical tests for determining the number of components. We also propose a way to reduce the number of predictors in a stepwise fashion, at each step eliminating the least important predictor based on a new measure of predictor importance. When at least one suppressor variable is included among the predictors, the new model predicts and validates better than traditional models, especially when G is large.

Keywords: K -component model, variable selection, gene expression, suppressor variable, latent class analysis

Acknowledgements: I am indebted to Karl Wassmann of Source MDx for his support and encouragement and for use of the melanoma data, to J. Alexander Ahlstrom for exceptional programming of CORExpress™, and Will Barker for his ongoing assistance. Multiple patent applications are pending regarding this technology.

1 Background and General approach

A new model with K correlated components is proposed for predicting an outcome variable Z based on G predictor variables. It is asymptotically equivalent to traditional regression models when $K = G$, and thus for $K < G$, it is more parsimonious than traditional regression models, despite the larger number of parameters. The approach to model development and a new method for variable reduction is illustrated on 2 data sets:

1. publicly available colon cancer vs. normals data containing $G = 2000$ continuous gene expression predictors - validated results show that a model based on only $G^* = 5$ genes provides strong prediction.
2. Survival data for 2 groups of melanoma patients. We begin with Z dichotomous and conclude by describing extensions where Z is ordinal, continuous, or nominal or Z indicating the occurrence of an event (e.g., death), or Z denotes multiple outcome variables.

Let Z denote an outcome variable and Y_1, Y_2, \dots, Y_G denote G continuous predictor variables. The K -component model is a generalized linear model (GLM), where the linear portion (ignoring the intercept) is obtained as a weighted sum of $K \leq G$ components S_1, S_2, \dots, S_K , each component itself being an exact linear combination of the predictors, $S_k = \sum_{g=1}^G \lambda_{kg} Y_g$. For concreteness, we will initially assume that Z is dichotomous, and denote the predicted logit obtained from the K -component model as $\text{Logit.K}(Z)$.

$$\text{Logit.K}(Z) = \alpha + \sum_{k=1}^K b_k^{(K)} S_k \quad (1)$$

$$= \alpha + \sum_{g=1}^G \beta_g^{(K)} Y_g \quad \text{where} \quad \beta_g^{(K)} = \sum_{k=1}^K b_k^{(K)} \lambda_{kg} \quad (2)$$

More generally, in a survival analysis a log-hazards rate might be used in the left-hand side of eq. (1) with a time varying intercept, or the conditional expectation of a continuous outcome variable as a linear regression extension. The approach to estimate the loadings, λ_{kg} , and weights, $b_k^{(K)}$, proceeds as follows: **Step 1:** Estimate loadings λ_{1g} defining the first component, $S_1 = \sum_{g=1}^G \lambda_{1g} Y_g$, using a maximum likelihood method, each term, $\lambda_{1g} Y_g$, corresponding to a GLM prediction of Z obtained from the g th predictor (omitting the intercept). When Z is dichotomous, each λ_{1g} corresponds to a simple log-odds ratio, the odds of $Z=1$ vs. $Z=0$ being $\exp(\lambda_{1g})$ times as high for a case having a 1 unit higher value on Y_g than another case. The CORE algorithm estimates the loadings, one at a time, as follows: Perform G separate linear regressions, the g th of which is the regression of Y_g on Z ,

$$Y_g = \alpha_{0g} + \lambda'_{0g} Z + \varepsilon_{0g} \quad (3)$$

Obtain an initial estimate for the loading λ_{1g} , denoted $\hat{\lambda}_{0g}$, by dividing the estimate for $\hat{\lambda}_{0g}$ by the mean squared error, $MSE(\varepsilon_{0g})$, obtained from the g th regression in (3):

$$\hat{\lambda}_{0g} = \hat{\lambda}'_{0g} / MSE(\varepsilon_{0g}) \quad (4)$$

Under the assumption that the error ε_{0g} is normally distributed with constant variance, $\hat{\lambda}_{0g}$ is a maximum likelihood estimate for the log-odds ratio λ_{0g} in the simple logistic regression model $\text{Logit}(Z|Y_g) = \alpha_g + \lambda_{0g} Y_g$ (Lyles, et. al. 2009). Using these G loadings, $\hat{\lambda}_{0g}$, $g=1,2,\dots,G$ the Naïve Bayes estimator for the component is $S_0 = \sum_{g=1}^G \hat{\lambda}_{0g} Y_g$, which is G times the average (or sum) of

G simple logistic regression model predictions (ignoring intercepts). The first component S_1 is then obtained as a standardized version of S_0 as follows: Perform a linear regression of S_0 on Z : $S_0 = \alpha_0 + b'_0 Z + \varepsilon_0$ Compute: $\lambda_{1g} = b_0 \lambda_{0g}$

and $S_1 = b_0 S_0$ where $b_0 = \hat{b}'_0 / MSE(\varepsilon_0)$. Predictions from the 1-component model are given by $Logit.1(Z) = \alpha + S_1$. Since $\beta_g^{(K)} = \sum_{k=1}^K b_k^{(K)} \lambda_{kg}$, the standardization of S_0 to S_1 is such that $b_1^{(1)} = 1$, which allows the gth loading on component S_1 to serve also as $\beta_g^{(1)}$, the weight for the gth predictor in the 1-component model. **Step 2:** Determine component S_2 such that it maximally improves prediction of Z over and above that provided by S_1 alone. The loadings on S_2 , denoted λ_{2g} , are estimated by a maximum likelihood method, where

$$S_2 = \sum_{g=1}^G \lambda_{2g} Y_g \quad (5)$$

The CORE algorithm proceeds as follows: Perform G separate linear regressions, the gth of which is the regression of Y_g on Z and S_1 , providing an estimate for λ'_{2g} in (6):

$$Y_g = \alpha_{2g} + \lambda'_{2g} Z + \gamma_1 S_1 + \varepsilon_{2g} \quad (6)$$

Also get the associated p-value testing the null hypothesis $H_0(1.g): \lambda'_{2g} = 0$, which serves as the equivalent test for the loading $\lambda_{2g} = 0$ where λ_{2g} in (5) is obtained by substituting the estimates for λ'_{2g} and MSE obtained in (6) into the equation $\lambda_{2g} = \lambda'_{2g} / MSE(\varepsilon_{2g})$. Obtain $Logit.2$ by estimating the b-weights in (1) corresponding to the logistic regression of Z on S_1 and S_2 :

$$Logit.2(Z|S_1, S_2) = \alpha + b_1^{(2)} S_1 + b_2^{(2)} S_2 \quad (7)$$

As in step 1 when we obtained $Logit.1$, we do not bother to estimate the intercept and the CORE algorithm obtains estimates for the b-coefficients as follows: Estimate the 2 linear regression models:

$$S_1 = a_1 + b'_{1.2} Z + d_1 S_2 + \varepsilon_1 \quad (8)$$

$$S_2 = a_2 + b'_{2.1} Z + d_2 S_1 + \varepsilon_2 \quad (9)$$

and compute:

$$b_1^{(2)} = b'_{1.2} / MSE(\varepsilon_1) \quad \text{and} \quad b_2^{(2)} = b'_{2.1} / MSE(\varepsilon_2) \quad (10)$$

From eq. (2) the composite weight for the gth constituent in the $K = 2$ component model is:

$$\beta_g^{(2)} = b_1^{(2)} \lambda_{1g} + b_2^{(2)} \lambda_{2g} \quad (11)$$

where λ_{1g} was obtained in Step 1, and λ_{2g} in Step 2. If the p-value associated with $H_0: b'_{2.1} = 0$ is non-significant, the 2nd component does not provide a significant improvement over the 1-component model, and the algorithm terminates with $K^*=1$. Otherwise, return to Step 2 with $K=K+1$. For example,

for $K=3$ determine component S_3 that improves prediction of Z over that provided by S_1 and S_2 alone. The algorithm terminates with the K^* -component model if the p-value associated with 1 or more $b_k^{(K^*+1)}$ is not statistically significant, in which case we say that the K^* -component model has achieved ‘sequential independence’.

Each predictor Y_g may have a different variance. From equation (2), reproduced here: ($Logit.K(Z) = \alpha + \sum_{g=1}^G \beta_g^{(K)} Y_g$), it is clear that standardizing Y_g by dividing by its standard deviation results in the associated composite weight being multiplied by the standard deviation. Thus, we define standardized composite weights $\beta_g^{*(K)} = \sigma_g \beta_g^{(K)}$, the absolute value of which we use as a measure of importance of variable g in the K -component model. Standardized loadings can be obtained by multiplying the corresponding raw loadings by the standard deviations: $\lambda_{kg}^* = \sigma_g \lambda_{kg}$.

We propose the following strategy to reduce the number of predictors from G to G^* : Given a value for K^* , eliminate the predictor variable with the lowest measure of importance $|\beta_g^{*(K^*)}|$, and re-estimate the 1-component through K^* -component models with $G-1$ predictors, determining again the lowest value for K for which sequential independence is achieved, and setting K^* to that value for K . Repeat this variable reduction process, eliminating 1 variable at a time until some stopping criteria is reached. For example, the stopping rule might be when a reduction in a validation performance measure occurs such as 1) AUC = the Area Under the ROC Curve, or 2) AMPS = Average Model Performance Statistic = $E(Logit.K|Z = 1) - E(Logit.K|Z = 0)$, as measured in validation data if available.

This new measure of predictor importance has advantages over other measures of importance. For a summary of weaknesses in measures of importance that have been proposed to date, see Gromping (2009).

Generally, S_2 is not predictive of Z directly, but is correlated with S_1 , and improves prediction by suppressing irrelevant variation in S_1 . In our analyses with gene expression data, we found that suppressor variables are prevalent and contribute most among all predictors in the model. For example, with respect to a 6-gene model for early detection of prostate cancer, the single most important variable was found to be a suppressor variable, on which the mean difference between the cancer and normal subjects was nil (Ross, et. al. 2010). For that model, the suppressor, called a ‘proxy gene’, enhances the predictive effects of 2 ‘prime genes’ in the model by predicting and subtracting out the expression value at an earlier time when the cancer subjects were normal, thus converting the gene expression for the prime genes to the more predictive variable representing the ‘change in expression’¹ on these prime genes.

¹ If component S_2 in eq. (7) is a pure suppressor, S_2 has no direct effect on Z and the linear regression of S_1 on S_2 has slope $m = -b_2/b_1$ so that $Logit.2(Z) = \alpha + b_1(S_1 - mS_2)$, S_2 enhancing the predictive power of S_1 .

Tables 1 and 2 show the results after application of the CORE variable selection algorithm to a training sample of $N=41$ cases to reduce the number of predictors from 2000 to 5 based on $K^*=4$. The goal is to discriminate between $Z=1$ ($N=40$ Colon cancer subjects) and $Z=0$ ($N=22$ Normal subjects). Table 2 shows that the predictor Hsa.25748 is a suppressor variable, since it does not load significantly on S_1 but has the sole significant loading on S_2 , which itself acts as a suppressor variable. Table 1 shows that Hsa.25748 is one of the most important variables in the K-component model, for $K = 3$ or 4. More generally, the powerful enhancement effects of suppressor variable is well documented (Friedman and Wall, 2005). Although common industry practice is to select from a large number of potential predictors only those that individually are predictive of the outcome variable(s), this strategy appears to be misguided, unnecessarily reducing the predictive power of a model by excluding proxy genes.

For the Colon Cancer Data, the 2-component model provides perfect prediction among the training data, and misclassifies only 3 in the validation data, 2 of which have been misclassified by many other models based on all 2000 genes. Results are similar for the 3-, 4-, and 5-component models. A larger number of misclassifications were reported from various PLS regression routines based on all 2000 genes (Fort, et. al., 2004).

Despite the fact that the 2-component model classifies all training subjects perfectly, the AMP statistic measured on the validation sample improves further when K increases from 2 to 3: $AMPS(1) = 8.5$, $AMPS(2) = 16.0$, $AMPS(3) = 17.4$, $AMPS(4) = 17.2$, the large improvement from $AMPS(1)$ being attributable largely to the enhancement effect of the suppressor variable. The p-values for the component weights (Table 2) show that inclusion of the 3rd component provides a marginally significant improvement ($p=.04$) and that the improvement for the 4th component is clearly non-significant ($p=.54$).

Table 1. Standardized composite weights (beta*) Results for suppressor variable (proxy gene) italicized

K	beta*(K)			
	1	2	3	4
Hsa.8125	-1.5	-3.3	-4.3	-4.7
Hsa.8147	-2.6	-4.4	-5.4	-4.9
Hsa.6814	1.3	2.6	2.2	2.2
Hsa.9353	1.0	2.2	1.9	2.0
Hsa.25748	<i>-0.2</i>	<i>3.0</i>	<i>4.1</i>	<i>4.0</i>
AMPS (Training, N=41)	7.3	13.4	15.1	14.9
AMPS (Validation, N=21)	8.5	16.0	17.4	17.2

Table 2. p-values for loadings and component weights (b)

k	p-values for loadings (lambda)			
	1	2	3	4
Hsa.6814	0.002	0.90	0.05	0.94
Hsa.8125	0.0007	0.57	0.05	0.54
Hsa.8147	2.1E-06	0.54	0.41	0.54
Hsa.9353	0.013	0.60	0.06	0.82
Hsa.25748	0.65	4.2E-05	0.64	0.54

k	p-values for component weights (b)			
	1	2	3	4
b(3,K)	6.9	7.2	5.9	
p	4.8E-14	1.8E-05	0.04	
b(4,K)	6.89	2.81	5.41	0.63
p	9.2E-14	0.01	0.03	0.54

The model generalizes easily in many ways. When Z is ordinal ² with known category scores, say 0, $Z^*[2], \dots, Z^*(J-1), 1$, the algorithm is unchanged, and for continuous Z only the division factor changes ³. For a second outcome variable ZB , or for Z nominal ⁴ (say, with $J = 3$ categories where $ZB = 1, 0$ is a second dummy indicator variable), ZB is included as an additional variable on the right side of eqs. (3) and (6), separate B-components, S_{1B}, S_{2B} , etc., are obtained and an additional equation (7B) for LogitB.K corresponding to eq. (7) with the additional components, along with corresponding additional eqs. (8B), (9B), (10B) and (11B). Extension to an ordinal Z with unknown category scores can also be obtained, the scores being estimated for each component k using a baseline logit model extension (see Magidson, 1996).

Another important generalization occurs when Z is a latent variable, being defined in an earlier analysis (step 0). For example, Z may be latent classes

² Without loss of generality any set of outcome scores $Z^*=Z^*[1], Z^*[2], \dots, Z^*[J]$ will be re-scaled to $Z=Z[1], Z[2], \dots, Z[J]$ such that $Z[1]=0$ and $Z[J]=1$, using: $Z[j] = (Z^*[j] - Z^*[1]) / (Z^*[J] - Z^*[1])$, $j = 1, 2, \dots, J$

³ An equivalent form of the algorithm uses Z as the left-hand side variable and Yg on the right and again uses least squares to estimate the parameter which is again divided by MSE. When Yg is continuous, the division by MSE is omitted; when the equivalent form is used with Z on the left-hand side variable, rather than division by MSE, the division is by the factor W which provides the same estimate as obtained with the original form. For example, when the variables are all standardized to have variance 1, when extracting S_2 , $W = (1 - r_{ZS_1}^2) / (1 - r_{Y_1S_1}^2)$. For Z continuous, an alternative is to use latent class analysis, predicting the high versus low scoring classes.

⁴ For $J > 3$, there would be $J-1$ additional dummy variables, say ZB, ZC, \dots

corresponding to subjects improving under a particular therapy (class 1) and those not improving (class 2), developed using multiple observed indicators of improvement. Example 2 below is based on 2 latent classes of melanoma patients, class 1 representing those who are ‘long term survivors’, and class 2 being ‘not long term survivors’, identified in a latent class survival analysis without any predictor variables. In this type of extension, the algorithm is unchanged if each case is assigned to a class based on their modal category, or it is extended to use weights corresponding to posterior membership probabilities obtained from the latent class model (see Magidson, 2005). The weights themselves may be case weights, or replication weights, or sampling weights where the case ID is used as a primary sampling unit using the Latent GOLD program (Vermunt and Magidson, 2008), the latter approach used below to get the appropriate p-values.

Previously, we obtained a 4-gene model that strongly predicted survival time among melanoma patients, which was found to validate when applied to a somewhat different melanoma population undergoing the same therapy (Bunkaitis-Davis, et. al, 2009). Here we re-analyzed that data using all G=169 genes, and used the selection algorithm to obtain a 4-gene model. We first used a latent class proportional hazards model to identify 2 classes - long term survivors and others - solely based on survival time (see Vermunt,2009), and obtained posterior membership probabilities for each class for each case. We then constructed a data file consisting of 2 records per case, one record for each class, along with 169 gene expression variables, the posterior membership probability as a weight. A weighted CORE analysis was then undertaken, replacing the OLS linear regression analyses with a weighted analysis. The resulting 4-gene model (Table 3) consists of 3 of the 4 genes obtained in the original analysis, the weakest (4th gene) in the original analysis being replaced by a different gene (identified as ‘gene X’). The resulting 4-gene model performed even better than the original model, on data from both populations, according to a log rank test where the risk score was used as a single covariate in a Cox model. Like the original model, the 4-genes again consisted of 2 prime and 2 ‘proxy genes’.

Table 3: Standardized composite weights (beta*) Results for suppressor variable (proxy gene) italicized

K	beta*(K)			
	1	2	3	4
Gene X	-0.20	-0.31	-0.20	-0.23
CTSD	-0.22	-0.62	-0.81	-0.82
PLA2G7	<i>0.06</i>	<i>0.41</i>	0.32	0.32
TXNRD1	<i>-0.01</i>	<i>0.49</i>	0.63	0.64

		p-values for component weights (b)			
k		1	2	3	4
b(3,K)		3.15	0.76	0.43	
p		1.4E-16	9.0E-13	0.05	
b(4,K)		2.16	1.28	0.41	0.32
p		1.0E-13	7.1E-13	0.05	0.96

		p-values for loadings (lambda)			
k		1	2	3	4
Gene X		2.0E-04	0.08	0.19	0.92
CTSD		4.4E-05	0.75	0.07	0.80
PLA2G7		<i>0.23</i>	<i>5.8E-04</i>	0.07	0.92
TXNRD1		<i>0.89</i>	<i>4.3E-08</i>	0.36	0.92

Table 4: p-values for component weights (b) and loadings Results for suppressor variable (proxy gene) italicized

References

- BUNKAITIS-DAVIS, D., F. GAO, J. MAGIDSON, K. WASSMANN (2010): Validation of a pre-treatment 4-gene Cox Model derived from peripheral blood gene expression measurements that predict survival in melanoma patients receiving CTLA4-blockade. *forthcoming*.
- FORT, GERSENDE and LAMBERT-LACROIX, SOPHIE. (2004): Classification Using Partial Least Squares with Penalized Logistic Regression. *IAP-Statistics*.
- FRIEDMAN, LYNN and WALL, MELANIE. (2005): Graphical Views Of Suppression And Muticollinearity In Multiple Linear Regression. *American Statistician*, May 2005. Vol 59, No. 2, pp 127-136.
- GROMPING, ULRIKE (2009): Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician*, November 2009. Vol 63, No. 4, pp 308-319.
- LYLES R.H., Y. GUO and A. HILL (2009): A Fresh Look at the Discrimination Function Approach for Estimating Crude or Adjusted Odds Ratios. *The American Statistician*, November 2009. Vol 63, No. 4, pp 320-327.
- MAGIDSON, JAY (1996): Maximum Likelihood Assessment of Clinical Trials Based on an Ordered Categorical Response. *Drug Information Journal*, Maple Glen, PA: Drug Information Association, Vol. 30, No. 1, 143-170.
- MAGIDSON, JAY (2005): An Extension of the CHAID Tree-based Segmentation Algorithm to Multiple Dependent Variables. in *C. Weihs & W. Gaul, Classification: The Ubiquitous Challenge*, 176-183. Heidelberg: Springer.
- ROSS, R.W., S. SENG, D. BANKAITIS-DAVIS, L. SICONOLFI, K. STORM, P. KANTOFF, J. MAGIDSON, K., WASSMANN, W. OH (2010): A Whole-Blood RNA Transcript-Based Diagnostic Test Improves the Diagnosis of Prostate Cancer Compared with Prostate-Specific Antigen Alone. *forthcoming*.
- VERMUNT, J.K. (2009): Event history analysis. in *R. Millsap and A. Maydeu-Olivares (eds.) Handbook of Quantitative Methods in Psychology*, 658-674. London: Sage.
- VERMUNT, J.K. and J. Magidson (2005): *Latent GOLD 4.0 Technical Guide*. Belmont MA.: Statistical Innovations Inc.