Some recent statistical learning methods for longitudinal high-dimensional data

Shuo Chen, University of Maryland
Edward Grant, Uniformed Services University
Tong Tong Wu, University of Rochester
Frederick Bowman, Emory University

Journal Title: Wiley Interdisciplinary Reviews: Computational Statistics
Volume: Volume 6, Number 1
Publisher: Wiley: 12 months | 2014-01-01, Pages 10-18
Type of Work: Article | Post-print: After Peer Review
Publisher DOI: 10.1002/wics.1282
Permanent URL: https://pid.emory.edu/ark:/25593/twbw7

Final published version: http://dx.doi.org/10.1002/wics.1282

Copyright information:
© 2013 Wiley Periodicals, Inc.

Accessed April 16, 2020 7:08 AM EDT
Statistical Learning Methods for Longitudinal High-dimensional Data

Shuo Chen¹, Edward Grant¹, Tong Tong Wu¹, and F. DuBois Bowman²
¹Department of Epidemiology and Biostatistics, University of Maryland, College Park, 20742
²Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322

Abstract

Recent studies have collected high-dimensional data longitudinally. Examples include brain images collected during different scanning sessions and time-course gene expression data. Because of the additional information learned from the temporal changes of the selected features, such longitudinal high-dimensional data, when incorporated with appropriate statistical learning techniques, are able to more accurately predict disease status or responses to a therapeutic treatment. In this article, we review recently proposed statistical learning methods dealing with longitudinal high-dimensional data.

Keywords

High-dimensionality; Multiple times points; Prediction; Support vector machines; Shrinkage; Temporal effects

1 Introduction

Current biomedical technology enables the collection of high-dimensional data longitudinally to gain understanding of genomic, proteomic, and in vivo neural processing properties over time. The temporal changes in high-profile biological properties may provide insight into disease diagnosis, progression, or recovery. Depending on the types of outcomes and clinical needs, the goals of longitudinal high-dimensional data include clustering and classification, survival analysis, multilevel regression and time series modeling.

By “longitudinal data,” we indicate two types of data collections: (1) high-dimensional profiles are collected at multiple times points during the study but the response variable is only collected at the end of the study as a final outcome; and (2) both the high-dimensional predictor variables and response variable are collected at multiple times points during the study. The desired methodology for high-dimensional longitudinal data would take advantage of the additional data to determine temporal trends of features and incorporate the temporal effects into learning methods and models that allow for repeated measurements. Recent research has developed several strategies to analyze high-dimensional longitudinal data using different statistical learning techniques, including support vector machines, non-parametric Bayesian methods, and shrinkage methods for different purposes. To address different objectives in the context of different data structures, we review several recent
methods for high-dimensional longitudinal data. Across these models, the key challenges are determining how to extract features in high-dimensional space and incorporate the temporal effects for more accurate prediction.

In this paper, we review a set of methods for high-dimensional longitudinal data, with focus on longitudinal support vector machine and penalized linear mixed effects models. We begin with basic concepts of each method, and then introduce how recent high-dimensional longitudinal data analysis methods extended from original model. We also review the computational strategies and algorithm implementations for these methods.

2 Methods

In this section, we will review several current statistical methods for use with both types of longitudinal high-dimensional data.

2.1 Longitudinal Support Vector Classifier - LSVC

We first review the statistical methods to deal with the first type of longitudinal high-dimensional data. The support vector (SVC) classifier is a robust and effective machine learning method that has been widely used for high-dimensional data analysis (Mitchell et al., 2004, Vapnik, 1996). Also, SVC has been applied to handle spatial-temporal high-dimensional data, Mourao-Miranda et al, 2007 first use singular value decomposition to obtain the linear combination of spatial and temporal effects and then apply the component as input of SVC. Recently, Chen and Bowman, 2011 developed a SVC based method for high-dimensional data measured at multiple time points. Suppose that longitudinal high-dimensional data is collected from \( N \) subjects at \( T \) measurement time points, and \( p \) represents the dimensionality of data. The expanded feature matrix for longitudinal high-dimensional data becomes \( TN \) by \( p \). For features \( x_{i,t} \) collected for subject \( i \) at time \( t \), the goal is to classify each individual \( x_i = \{x_{i,1}, x_{i,2}, ..., x_{i,T}\}' \) to certain groups \( y_i \in \{-1, 1\} \), outcomes only collected at the end of the study.

Linear trends of change are characterized: \( x_s = x_{i,1} + \beta_1 x_{i,2} + \beta_2 x_{i,3} ... + \beta_{T-1} x_{i,T} \), where \( \beta = (1, \beta_1, \beta_2, ..., \beta_{T-1})' \) is an unknown parameter vector. Such trend information is highly desired to improve the classification accuracy, usually not available. Thus, a key challenge of building a classifier of longitudinal high dimensional data is jointly estimating the separating hyperplane parameters \( a \) and the temporal trend parameters \( \beta \). Chen and Bowman, 2011 first proposed a novel longitudinal support vector classifier (LSVC) to solve the problem using quadratic programming.

The LSVC is extended from the conventional support vector classifier (SVC), and it augments the cross-sectional high-dimensional feature space to a longitudinal high-dimensional feature space. The method seeks to construct an objective function by incorporating both temporal trend parameters and separating the hyperplane parameters. In the paper, the authors first note the augmented Gram matrix as
\[ G = (\tilde{X}_{t=1} + \beta_1 \tilde{X}_{t=2} + \ldots + \beta_{T-1} \tilde{X}_{t=T})^T (\tilde{X}_{t=1} + \beta_1 \tilde{X}_{t=2} + \ldots + \beta_{T-1} \tilde{X}_{t=T}) = (\tilde{X}_m \beta_m)^T (\tilde{X}_m \beta_m) = \beta_m^T G_m \beta_m, \]

where

\[
G_m = \begin{bmatrix}
\tilde{X}_{t=1}^T \tilde{X}_{t=1} & \cdots & \tilde{X}_{t=1}^T \tilde{X}_{t=T} \\
\vdots & \ddots & \vdots \\
\tilde{X}_{t=T}^T \tilde{X}_{t=1} & \cdots & \tilde{X}_{t=T}^T \tilde{X}_{t=T}
\end{bmatrix},
\]

\[ \tilde{X}_m = [\tilde{X}_{t=1}, \tilde{X}_{t=2}, \ldots, \tilde{X}_{t=T}]^T \]
represents the \( p \times TN \) longitudinal high dimensional features, with components \( \tilde{X}_{t=k} = (y_1 x_{1,t=k}, y_2 x_{2,t=k}, \ldots, y_N x_{N,t=k}) \) to be data from \( N \) subjects each with \( p \) features at time point \( k \). The corresponding \( \beta_m \) is a \( TN \times N \) matrix, and \( \beta_m^T = [I_{N \times N}, \beta_1 I_{N \times N}, \beta_2 I_{N \times N}, \ldots, \beta_{T-1} I_{N \times N}] \). Similar to the conventional SVC, the objective function of LSVC is also subject to maximize the margins in the following equation:

\[
\min \frac{1}{2} \| w_{nv} \|^2 + \sum_{i=1}^{N} \xi_i, \quad \text{for } i = 1, \ldots, N \tag{2.1}
\]

where \( w_{nv} \) is the estimate of separating hyperplane parameters with, by assuming that the temporal trend parameters are known, longitudinal high-dimensional features \( x_i = x_{i,1} + \beta_1 x_{i,2} + \beta_2 x_{i,3} + \ldots + \beta_{T-1} x_{i,T} \).

After the incorporation of the temporal trend parameters \( w_{nv} = \sum_{s=1}^{N} y_s \alpha_s (\tilde{x}_s \beta_m)^T \), the Langrange (Wolfe) dual function becomes:

\[
\min \frac{1}{2} \alpha_m^T G_m \alpha_m - 1^T \alpha \\
\text{subject to} \\
C \geq \alpha_m(i) \geq 0, \\
\sum_{t=1}^{T} \sum_{i=1}^{N} \alpha_m[i+(t-1)N] y_i = 0, 
\tag{2.2}
\]

for \( i = 1, \ldots, N \) and \( t = 1, \ldots, T - 1 \).

Provided with \( \alpha_m \), the separating hyperplane parameter becomes

\[
w_{nv} = \sum_{i=1}^{n} y_i \alpha_m(i)(x_{i,1} + \beta_1 x_{i,2} + \beta_2 x_{i,3} + \ldots + \beta_{T-1} x_{i,T}),
\]

where \( \alpha_m = (\alpha_m(i), \alpha_m(i+N), \ldots, \alpha_m(i+(T-1)N)) \) and \( w_{nv} = \sum_{i=1}^{n} y_i \alpha_m(i) \tilde{x}_i \).
Given $w_{m}$, the intercept term then becomes

$$b = \frac{1}{N} \sum_{i=1}^{N} (w_{m} \cdot (x_{i} \beta_{m})^{T} - y_{i})$$

in which $\beta_{m}$ can be estimated based on $\alpha_{m}$. At last, the separating hyperplane can be used to classify each subject by

$$h(x) = w_{m} \cdot (x \beta_{m})^{T} + b.$$  \hspace{1cm} (2.3)

**Model Estimation**—To estimate $\alpha$ and $\beta$ vectors, the authors suggest to reparameterize the first part of the objective function in 2.6 as:

$$f = \left( \begin{array}{c} \alpha \\ \beta \alpha \end{array} \right)^{T} \left[ \begin{array}{cc} G_{m}^{0,0} & G_{m}^{0,1} \\ G_{m}^{1,0} & G_{m}^{1,1} \end{array} \right] \left( \begin{array}{c} \alpha \\ \beta \alpha \end{array} \right)$$

where

$$G_{m} = \left[ \begin{array}{cc} G_{m}^{0,0} & G_{m}^{0,T} \\ G_{m}^{T,0} & G_{m}^{T,T} \end{array} \right],$$

and $G_{m}^{0,0}$ is the $N \times N$ submatrix in the left top corner of the matrix $G_{m}$ for the baseline data ($\tilde{X}_{T-1}$), the Gram matrix of SVC.

The objective function has been proven to be convex, and an iterative quadratic programming (QP) procedure is developed for optimization: (1) start with initial values of $\beta$ and use QP to optimize 2.3 to obtain $\alpha$; (2) use the updated $\alpha$ obtained in step 1 and apply QP again to estimate $\beta$; and (3) repeat the above two steps until convergence. The convergence of the iterative algorithm can be achieved because a unique solution exists.

**Nonlinear Kernel Functions**—The authors also provide solutions for nonlinear kernels. The Gram matrix of a nonlinear kernel is

$$K(\tilde{x}_{i}, \tilde{x}_{j}^{'}_{T}) = \left[ \begin{array}{ccc} K(\tilde{x}_{i,1}, \tilde{x}_{j,1}^{'}_{T}) & \cdots & K(\tilde{x}_{i,1}, \tilde{x}_{j,T}^{'}_{T}) \\ \vdots & \ddots & \vdots \\ K(\tilde{x}_{i,T}, \tilde{x}_{j,1}^{'}_{T}) & \cdots & K(\tilde{x}_{i,T}, \tilde{x}_{j,T}^{'}_{T}) \end{array} \right],$$

where $\langle K(\cdot, x_{j}^{'}_{T}), K(\cdot, x_{j}^{'}_{T}) \rangle = K(x_{j}^{'}_{T}, x_{j}^{'}_{T})$, and $K(\cdot, x_{j}^{'}_{T})$ indicates the reproducing kernel map of $x_{j}^{'}_{T}$ (Wahba, 1990). The separating hyperplane with a nonlinear kernel becomes
where $b$ is obtained by 

$$b = \frac{1}{N} \sum_{i=1}^{N} \sum_{i'=1}^{N} (y_i y'_i \alpha m \bar{K}(\bar{x}_i, \bar{x}_i') \beta_m^T - y_i).$$

Therefore, the nonlinear kernel does not increase the complexity of estimating $\beta$. In addition, the authors discussed the variable selection based on the predictors’ effect on the objective function which is a “wrapper” method (Guyon et al., 2003; Hastie and Tibshirani, 2004).

To demonstrate the use and potential advantages of LSVC, the authors apply the method to a simulation study and a data example from the Alzheimer’s disease Neuroimaging Initiative. The results show that by leveraging the additional longitudinal information LSVC achieves higher accuracy than methods using only cross-sectional data and methods that combine longitudinal data by naively expanding the feature space.

### 2.2 Penalized Linear Mixed Effects Models

Linear mixed effects models can be used in the analysis of clustered or longitudinal data. Those models estimate the relationship between the dependent variable and the fixed effects and random effects of independent variables by considering both means and covariances. With the improvement of data collection and storage technology, a large number of independent variables are available and can be included in the model. Inference and prediction of such a model becomes too complex and infeasible when the number of predictors increases. One challenge is how to choose significant predictors while excluding variables that have no true effects on the outcome. An example is the Trial of Activity for Adolescent Girls (TAAG) study, which determined the effectiveness of a school and community based intervention on the physical activities of girls from 6 middle schools in Maryland (Young et al., 2013). A large group of girls were followed up for four years from year 2006 to 2009 and asked to take a survey with hundreds of questions at multiple time points to measure the change of physical activities. A linear mixed effects model can be fitted to take into account the clustering effects (6 middle schools) and temporal effects (four years) as well as the fixed effects such as race, socio-economic status, and other questions in the survey.

To construct a linear mixed effects model, consider the $i$th subject in a longitudinal study with $n$ subjects, each having observations at $m_i$ time points for a total of $N = \sum_{i=1}^{n} m_i$ observations. The linear mixed effects model can be written by

$$y_{ij} = x_{ij}^{T}\beta + z_{ij}^{T}b_i + \varepsilon_{ij}, \quad j=1, \ldots, m_i$$

where $y_{ij}$ is the response variable at the $j$th time point, $x_{ij}$ is the vector of $p$ fixed effects, $z_{ij}$ is the vectors of $q$ random effects at the $j$th time point, $\beta$ is the $p$ parameter vector for the fixed effects, $b_i$ is the $q$ parameter vector for random effects, and $\varepsilon_{ij}$ is the i.i.d. random error from
The random effects parameter $b_i$ are i.i.d. multivariate normal variables following $MVN(0, \sigma^2 \Pi)$, where $\Pi$ is the covariance matrix, and are independent of $\epsilon_{ij}$. Using the matrix notation, the model can be simplified as

$$Y_i = X_i \beta + Z_i b_i + \epsilon_i,$$  \hspace{75pt} (2.4)$$

where $Y_i = (y_{i1}, …, y_{imi})^T$, $X_i = (x_{i1}, …, x_{imi})^T$ is the $m \times p$ design matrix of fixed effects, $Z_i = (z_{i1}, …, z_{imi})^T$ is the $m \times q$ design matrix of random effects, and $\epsilon_i = (\epsilon_{i1}, …, \epsilon_{imi})^T$ is an i.i.d. random error following $N(0, \sigma^2 I_m)$.

Previous methods for penalized estimation of fixed effects include Efron et al., 2004, Zou and Hastie, 2005, and Bondell and Reich, 2008. Previous methods for selection of random effects include Stram and Lee, 1994, Lin, 1997, Hall and Praestgaard, 2001, and Chen, 2003. In this review we focus on three penalized linear mixed effects models, which, unlike many previous approaches, select both fixed effects and random effects simultaneously. The first model is developed by Bondell et al., 2010, maximizing a penalized joint likelihood problem. The second model is introduced by Fan and Li, 2012, uses a proxy matrix in maximizing a penalized profile likelihood for fixed and random effects separately. Third, the paper of Li et al., 2012 optimizes maximum likelihood estimator with separate penalization methods for fixed and random effects. A side-by-side comparison of these models is summarized in Table 1.

**Model 1 (Bondell et al., 2010)**—Method 1 estimates fixed effects, random effects, and the covariance structure of the selected random effects simultaneously in a model with one penalty function. Equation (2.4) can be reparameterized using a modified Cholesky decomposition to factorize the covariance matrix of the random effects, $\Pi$. Through this factorization, $\Pi = D \Gamma \Gamma^T D$, where $\Gamma$ is a $q \times q$ lower triangular matrix with 1’s on the diagonal and whose $(l, r)$th element is given by $\gamma_{lr}$, and $D = \text{diag}(d_1, d_2, …, d_q)$ is a diagonal matrix. After this reparameterization, the linear mixed effects model (2.4) becomes:

$$Y_i = X_i \beta + Z_i D \Gamma b_i + \epsilon_i.$$

The covariance matrix of $b_i$ is now expressed in terms of vector $d = (d_1, d_2, …, d_q)^T$ and the free elements of $\Gamma$, denoted by vector $\gamma = (\gamma_{lr} : l = 1, …, q : r = l + 1, …, q)^T$. Setting any $d_l = 0$ will set the corresponding $l$th row and column of the covariance matrix $\Pi$ to 0 and therefore remove the $l$th random effect from the model.

After reparameterizing the model and treating $b$ as given, maximizing the log-likelihood function is equivalent to minimizing the conditional expectation of $||y - Z D \tilde{\Gamma} \tilde{b} - X \beta ||^2$, where $D = I_m \otimes D$ and $\Gamma = I_m \otimes \Gamma$. Rearranging the terms and adding the Adaptive Least Absolute Selection and Shrinkage Operator (ALASSO, Zou, 2006) penalty, the goal is to minimize the quadratic problem:
\[ Q(\beta^T, \alpha^T, \gamma^T | y, b) = \| y - X\beta - Z\text{diag}(\Gamma b)(1_q \otimes I_m) \|^2 + \lambda \left( \sum_{j=1}^p |\beta_j| + \sum_{k=1}^q |d_k| \right), \quad (2.5) \]

where $\hat{\beta}_j$ and $d_k$ are ordinary least squares estimates and $1_q$ is a column vector of ones of length $q$. The problem (2.5) can be solved using the EM algorithm developed by Laird and Ware (1982) and Laird, Lange, and Stram (1987).

For high-dimensional data, it would be necessary to reduce the dimension of the data before using the method. This could be accomplished by using previous methods of penalized variable selection on the fixed effects, while ignoring the random effects, and vice versa. The resulting reduced model could then be applied to this joint penalty problem for further simultaneous selection of fixed and random effects. While this method is effective at selecting fixed and random effects, the EM algorithm that it uses is not efficient and may not be plausible when the number of predictors is too large due to its slow convergence rate and computational burden.

**Model 2 (Fan and Li, 2012)**—This method selects important fixed and random effects independently in two separate models. Proxy matrices are used to account for the unknown variance-covariance structure of the random effects during the selections. Stacking $X_i, b_i, y_i,$ and $\epsilon_i$ and setting $Z = \text{diag}(Z_1, \ldots, Z_n)$ with corresponding $\Pi = \text{diag}(\Pi, \ldots, \Pi)$, the linear mixed effects model in (2.4) can be rewritten as

\[ y = X\beta + Zb + \epsilon. \]

For the fixed effects parameter $\beta$, it is necessary to minimize the penalized likelihood equation

\[ Q(\beta) = \frac{1}{2} (y - X\beta)^T P_{\sigma} (y - X\beta) + n \sum_{j=1}^p P(\|\beta_j\|), \quad (2.6) \]

where $P_{\sigma} = (I + \sigma^2 Z\Pi Z^T)^{-1}$ and the penalty function $P(\|\beta_j\|)$ is the Smoothed Clipped Absolute Deviation (SCAD, Fan and Li, 2001) penalty with tuning parameter $\lambda$. It is important to note that the problem $Q(\beta)$ depends on the unknown parameters $\Pi$ and $\sigma^2$. To overcome this obstacle, a proxy matrix $P_{\tilde{\sigma}} = (I + Z M Z^T)^{-1}$, where $M = (\log n) I$, is substituted into (2.6) for $P_{\sigma}$. Since this regularization function is quadratic, it can be solved through previous methods for penalized least-squares, such as the LARS algorithm (Efron et al., 2004).

The selection of the random effects is accomplished through Bayesian methods of deriving the restricted posterior distribution of the random effects, and penalizing this solution of the restricted posterior mode. The resulting regularization problem to be minimized is
\[
Q(b) = \frac{1}{2}(y - Zb)^T P_x(y - Zb) + \frac{1}{2}\sigma^2 b^T \tilde{\Pi} b + n \sum_{k=1}^{q} P_{\lambda}(|b_k|), \quad (2.7)
\]

where \( P_x = I - X(X^TX)^{-1}X^T \), the penalty function \( P_{\lambda}(|b_k|) \) is again the SCAD penalty with tuning parameter \( \lambda \), and \( \tilde{\Pi} \) is the Moore-Penrose generalized inverse of \( \Pi \). Again, \( \Pi \) and \( \sigma^2 \) are unknown so a proxy matrix \( \mathcal{M} = \text{diag}(M, \ldots, M) \), with \( M = (\log n) \), is substituted for \( \sigma^{-2} \Pi \), so the regularization problem in (2.7) becomes:

\[
Q(b) = \frac{1}{2}(y - Zb)^T P_x(y - Zb) + \frac{1}{2} b^T \mathcal{M} b + n \sum_{k=1}^{q} P_{\lambda}(|b_k|).
\]

This problem is similar to the penalized quadratic function of adaptive elastic net (Zou and Zhang, 2009), so it can be solved through modification of this algorithm.

For high-dimensional data where \( N \leq p \), the dimension of fixed effects must be lowered to below the sample size before using the above methods. This can be done by first using penalized least squares methods on the fixed effects while ignoring random effects. Using these selected fixed effects, the random effects can be estimated using the regularization problem (2.7). Next, using these selected random effects from the second step, the fixed effects regularization problem (2.6) can be used to select from the remaining fixed effects. The second and third steps can be repeated, as needed, to further reduce the dimensionality of the data.

**Model 3 (Li et al., 2012)—** The final method selects and estimates fixed effects, random effects, and the covariance structure of the selected random effects simultaneously in a linear mixed effects model using two penalty functions. Using the model (2.4), When \( N > p \), a modified log-likelihood incorporating the Restricted Maximum Likelihood (REML) is

\[
\ell_{nM}(\beta, \Pi, \sigma^2) = -\frac{1}{2} \sum_{i=1}^{n} \log |\sigma^2 V_i| - \frac{1}{2} \log \sigma^2 - \frac{1}{2} \sum_{i=1}^{n} X_i^T V_i^{-1} X_i - \frac{1}{2 \sigma^2} \sum_{i=1}^{n} (Y_i - X_i \beta)^T V_i^{-1} (Y_i - X_i \beta), \quad (2.8)
\]

where \( V_i = I_{n_i} + Z_i \Pi Z_i^T \). In high dimensional settings when \( N \leq p \), the restricted term in (2.8) will become singular so the following full log-likelihood must be used

\[
\ell_{nF}(\beta, \Pi, \sigma^2) = -\frac{1}{2} \sum_{i=1}^{n} \log |\sigma^2 V_i| - \frac{1}{2 \sigma^2} \sum_{i=1}^{n} (Y_i - X_i \beta)^T V_i^{-1} (Y_i - X_i \beta). \quad (2.9)
\]

The maximum likelihood can be found for this equation to obtain the parameters. Adding penalty functions, the regularization problem to maximize is

\[
Q_n(\beta, \Pi, \sigma^2) = \ell_n(\beta, \Pi, \sigma^2) - \lambda_1 P_1(\beta) - \lambda_2 P_2(\mathbf{D}), \quad (2.10)
\]
where $\ell_n(\theta)$ is (2.8) or (2.9) depending on if $N > p$ or $N \leq p$, respectively. The first penalty function, $\lambda_1 P_1(\beta)$, is for the fixed effects and is an ALASSO penalty. For the random effects, a Cholesky decomposition is performed such that $\Pi = LL^T$ where $L$ is a lower triangular matrix with positive diagonal elements. If any diagonal element of $L_{kk} = 0$, then the corresponding random effect $b_k$ is also 0 and is removed from the model. The second penalty function in (2.10) is set to be an $L - 2$ penalty function (Yuan and Lin, 2006) with an adaptive weight added so that the regularization problem is

$$Q_n(\beta, \Pi, \sigma^2) = \ell_n(\beta, \Pi, \sigma^2) - \lambda_1 \sum_{j=1}^{p} \sum_{k=2}^{q} \sqrt{L_{k1}^2 + \ldots + L_{kq}^2},$$

(2.11)

where $\hat{\beta}_j$ and $\|L_{(k)}\|$ are ordinary least squares estimates. An estimation of the variance $\sigma^2$ (Lindstrom and Bates, 1988) can be substituted into the model, allowing (2.11) to be maximized in terms of only $L$ and $\beta$.

The problem (2.11) can be solved by a new algorithm that iteratively updates two quadratic optimization functions for the random and fixed effects. This has proven to be more efficient than the EM algorithm, which cannot handle large numbers of predictors. When the maximum likelihood is used, the new algorithm is proven to be a consistent estimator for high-dimensional data, where $p$ and $q$ can diverge at an exponential rate with the sample size $n$.

3 Summary

In this article, we described two types of longitudinal high-dimensional data that researchers often encounter in current biomedical research and reviewed several recently developed statistical methods to deal with these two types of data. First, we introduced a kernel method for classification for the first type of longitudinal high-dimensional data and the corresponding computational strategy for parameter estimation. Second, we reviewed three mixed effect shrinkage models for the other type of longitudinal high-dimensional data. In the review, we compared the model setups, computational strategies, and advantages and shortcomings of the methods.

Acknowledgments

Wu’s research is supported in part by NSF Grant CCF-0926181.

References


Vapnik, V. The nature of statistical learning theory. New York: Springer; 1996. p. 188


## Table 1

Comparison of three penalized linear mixed effects models

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Regularizations</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Objective Functions</td>
<td>$\ell(\beta, \Pi, \sigma^2) - \lambda P(\beta, b)$</td>
<td>Fixed: $\ell(\beta) - \lambda_1 P_1(\beta)$ Random: $\ell(b, \sigma^2) \lambda_2 P_2(b)$</td>
<td>$\ell(\beta, \Pi, \sigma^2) - \lambda_1 P_1(\beta) - \lambda_2 P_2(b)$</td>
</tr>
<tr>
<td>Penalties</td>
<td>1 total (ALASSO)</td>
<td>1 fixed (SCAD)</td>
<td>1 fixed (ALASSO)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 random (SCAD)</td>
<td>1 random (L-2 norm)</td>
</tr>
<tr>
<td>Covariance Structure</td>
<td>Modified Cholesky decomposition for covariance matrix</td>
<td>Use of proxy matrix to substitute for unknown covariance matrix</td>
<td>Cholesky decomposition for covariance matrix</td>
</tr>
<tr>
<td>Algorithms</td>
<td>EM algorithm</td>
<td>LARS/Elastic net</td>
<td>New efficient algorithm with two quadratic components</td>
</tr>
<tr>
<td>High-Dimensional Data</td>
<td>EM algorithm is not efficient for large number of predictors</td>
<td>$p$ and $q$ can diverge to $\infty$, but must reduce dimensions of fixed effects ignoring random effects first</td>
<td>$p$ and $q$ can diverge to $1$</td>
</tr>
</tbody>
</table>