Published in final edited form as:
Wiley Interdiscip Rev Comput Stat. 2014 ; 6(1): 10–18. doi:10.1002/wics.1282.

Statistical Learning Methods for Longitudinal High-dimensional Data

Shuo Chen[1], **Edward Grant**[1], **Tong Tong Wu**[1], and **F. DuBois Bowman**[2]

[1]Department of Epidemiology and Biostatistics, University of Maryland, College Park, 20742

[2]Department 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 $\boldsymbol{x}_{i,t}$ collected for subject i at time t, the goal is to classify each individual $\tilde{\boldsymbol{x}_i} = \{\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \ldots, \boldsymbol{x}_{i,T}\}'$ to certain groups $y_i \in \{-1, 1\}$, outcomes only collected at the end of the study.

Linear trends of change are characterized: $\boldsymbol{x}_s = \boldsymbol{x}_{i,1} + \beta_1\boldsymbol{x}_{i,2} + \beta_2\boldsymbol{x}_{i,3}\ldots + \beta_{T-1}\boldsymbol{x}_{i,T}$, where $\boldsymbol{\beta} = (1, \beta_1, \beta_2, \ldots, \beta_{T-1})'$ 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 $\boldsymbol{\alpha}$ and the temporal trend parameters $\boldsymbol{\beta}$. 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

$$\begin{aligned}\mathbf{G} &= (\tilde{\mathbf{X}}_{t=1}+\boldsymbol{\beta}_1\tilde{\mathbf{X}}_{t=2}+\ldots+\boldsymbol{\beta}_{T-1}\tilde{\mathbf{X}}_{t=T})^T(\tilde{\mathbf{X}}_{t=1}+\boldsymbol{\beta}_1\tilde{\mathbf{X}}_{t=2}+\ldots+\boldsymbol{\beta}_{T-1}\tilde{\mathbf{X}}_{t=T})\\ &= (\tilde{\mathbf{X}}_m\boldsymbol{\beta}_m)^T(\tilde{\mathbf{X}}_m\boldsymbol{\beta}_m)\\ &= \boldsymbol{\beta}_m^T\mathbf{G}_m\boldsymbol{\beta}_m,\end{aligned}$$

where

$$\mathbf{G}_m=\begin{bmatrix} \tilde{\mathbf{X}}_{t=1}^T\tilde{\mathbf{X}}_{t=1} & \cdots & \tilde{\mathbf{X}}_{t=1}^T\tilde{\mathbf{X}}_{t=T}\\ \vdots & \ddots & \vdots\\ \tilde{\mathbf{X}}_{t=T}^T\tilde{\mathbf{X}}_{t=1} & \cdots & \tilde{\mathbf{X}}_{t=T}^T\tilde{\mathbf{X}}_{t=T}\end{bmatrix},$$

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

$$\min_{\boldsymbol{w}_{nv}}\frac{1}{2}\|\boldsymbol{w}_{nv}\|^2+\mathscr{C}\sum_{i=1}^{N}\xi_i, \quad \text{for } i=1,\ldots,N \qquad (2.1)$$

where $\boldsymbol{w}_{nv}$ is the estimate of separating hyperplane parameters with, by assuming that the temporal trend parameters are known, longitudinal high-dimensional features $\boldsymbol{x}_i = \boldsymbol{x}_{i,1} + \beta_1\boldsymbol{x}_{i,2} + \beta_2\boldsymbol{x}_{i,3}\ldots + \beta_{T-1}\boldsymbol{x}_{i,T}$.

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

$$\begin{aligned} &\min_{\boldsymbol{\alpha}}\tfrac{1}{2}\boldsymbol{\alpha}_m^T\mathbf{G}_m\boldsymbol{\alpha}_m-\mathbf{1}^T\boldsymbol{\alpha}\\ \text{subject to}\quad &C \geq \boldsymbol{\alpha}_m(i) \geq 0,\\ &\sum_t^T\sum_i^N\boldsymbol{\alpha}_m[i+(t-1)N]y_i=0,\end{aligned} \qquad (2.2)$$

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

Provided with α_m, the separating hyperplane parameter becomes

$$\boldsymbol{w}_{nv}=\sum_{i=1}^{n}y_i\boldsymbol{\alpha}_m(i)(\boldsymbol{x}_{i,1}+\beta_1\boldsymbol{x}_{i,2}+\beta_2\boldsymbol{x}_{i,3}\ldots+\beta_{T-1}\boldsymbol{x}_{i,T}),$$

where $\alpha_{m,i} = (\alpha_m(i), \alpha_m(i + N), \ldots, \alpha_m(i + (T - 1)N))$ and $\boldsymbol{w}_{nv}=\sum_{i=1}^{n}y_i\boldsymbol{\alpha}_{m,i}\tilde{\boldsymbol{x}}_i$.

Given $\boldsymbol{w}_{nv}$, the intercept term then becomes $b=\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{w}_{nv}\cdot(\tilde{\boldsymbol{x}}_i\boldsymbol{\beta}_m)^T-y_i)$, in which β_m can be estimated based on α_m. At last, the separating hyperplane can be used to classify each subject by

$$h(\tilde{\boldsymbol{x}})=\boldsymbol{w}_{nv}\cdot(\tilde{\boldsymbol{x}}\boldsymbol{\beta}_m)^T+b. \quad (2.3)$$

Model Estimation—To estimate α and β vectors, the authors suggest to reparameterize the first part of the objective function in 2.6 as:

$$f=\begin{pmatrix}\boldsymbol{\alpha}\\ \beta\boldsymbol{\alpha}\end{pmatrix}^T\begin{bmatrix}\mathbf{G}_m^{0,0} & \mathbf{G}_m^{0,1}\\ \mathbf{G}_m^{1,0} & \mathbf{G}_m^{1,1}\end{bmatrix}\begin{pmatrix}\boldsymbol{\alpha}\\ \beta\boldsymbol{\alpha}\end{pmatrix}$$

where

$$\mathbf{G}_m=\begin{bmatrix}\mathbf{G}_m^{0,0} & \mathbf{G}_m^{0,T}\\ \mathbf{G}_m^{T,0} & \mathbf{G}_m^{T,T}\end{bmatrix},$$

and $\mathbf{G}_m^{0,0}$ is the $N \times N$ submatrix in the left top corner of the matrix $\mathbf{G}_m$ for the baseline data ($\tilde{\mathbf{X}}_{t=1}^T\tilde{\mathbf{X}}_{t-1}$), 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 β and use QP to optimize 2.3 to obtain α; (2) use the updated α obtained in step 1 and apply QP again to estimate β; 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

$$\tilde{\mathbf{K}}(\tilde{\boldsymbol{x}}_i,\tilde{\boldsymbol{x}}_{i'})=\begin{bmatrix}\mathbf{K}(\tilde{\boldsymbol{x}}_{i,1},\tilde{\boldsymbol{x}}_{i',1}) & \cdots & \mathbf{K}(\tilde{\boldsymbol{x}}_{i,1},\tilde{\boldsymbol{x}}_{i',T})\\ \vdots & \ddots & \vdots\\ \mathbf{K}(\tilde{\boldsymbol{x}}_{i,T},\tilde{\boldsymbol{x}}_{i',1}) & \cdots & \mathbf{K}(\tilde{\boldsymbol{x}}_{i,T},\tilde{\boldsymbol{x}}_{i',T})\end{bmatrix},$$

where $\langle\beta\mathbf{K}(\cdot,\boldsymbol{x}_{i,\tilde{t}}),\mathbf{K}(\cdot,\boldsymbol{x}_{i',\tilde{t}})\rangle=\beta\mathbf{K}(\boldsymbol{x}_{i,\tilde{t}},\boldsymbol{x}_{i',\tilde{t}})$, and $\mathbf{K}(\cdot,\boldsymbol{x}_{i,\tilde{t}})$ indicates the reproducing kernel map of $\boldsymbol{x}_{i,\tilde{t}}$ (Wahba, 1990). The separating hyperplane with a nonlinear kernel becomes

$$h(\tilde{\boldsymbol{x}})=\sum_{s=1}^{N} y_s \boldsymbol{\alpha}_m \tilde{\mathbf{K}}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}_s)\boldsymbol{\beta}_m^T+b,$$

where b is obtained by $b=\frac{1}{N}\sum_{i=1}^{N}\sum_{i'=1}^{N}(y_i y_i^{'} \boldsymbol{\alpha}_m \tilde{\mathbf{K}}(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_{i'})\boldsymbol{\beta}_m^T-y_i)$. Therefore, the nonlinear kernel does not increase the complexity of estimating $\boldsymbol{\beta}$. 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 ith 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

$$y_{ij}=\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\boldsymbol{z}_{ij}^T\boldsymbol{b}_i+\varepsilon_{ij}, \quad j=1,\ldots,m_i$$

where y_{ij} is the response variable at the jth time point, $\boldsymbol{x}_{ij}$ is the vector of p fixed effects, $\boldsymbol{z}_{ij}$ is the vectors of q random effects at the jth time point, $\boldsymbol{\beta}$ is the p parameter vector for the fixed effects, $\boldsymbol{b}_i$ is the q parameter vector for random effects, and ε_{ij} is the i.i.d. random error from

$N(0, \sigma^2)$. The random effects parameter $\boldsymbol{b}_i$ are i.i.d. multivariate normal variables following $MVN(0, \sigma^2\Pi)$, where Π is the covariance matrix, and are independent of ε_{ij}. Using the matrix notation, the model can be simplified as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \quad (2.4)$$

where $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{im_i})^T$, $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{im_i})^T$ is the $m \times p$ design matrix of fixed effects, $\boldsymbol{Z}_i = (z_{i1}, \ldots, z_{im_i})^T$ is the $m \times q$ design matrix of random effects, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im_i})^T$ is an i.i.d. random error following $N(0, \sigma^2\boldsymbol{I}_{m_i})$.

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, Π. Through this factorization, $\Pi = \boldsymbol{D}\Gamma\Gamma^{\boldsymbol{T}}\boldsymbol{D}$, where Γ is a $q \times q$ lower triangular matrix with 1's on the diagonal and whose (l, r)th element is given by γ_{lr} and $\boldsymbol{D} = \mathrm{diag}(d_1, d_2, \ldots, d_q)$ is a diagonal matrix. After this reparameterization, the linear mixed effects model (2.4) becomes:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{\Gamma}\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i.$$

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

After reparameterizing the model and treating $\boldsymbol{b}$ as given, maximizing the log-likelihood function is equivalent to minimizing the conditional expectation of $\|\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{D}}\tilde{\Gamma}\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta}\|^2$, where $\tilde{\boldsymbol{D}} = I_m \otimes \boldsymbol{D}$ and $\tilde{\Gamma} = I_m \otimes \Gamma$. 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:

$$Q(\boldsymbol{\beta}^T, \boldsymbol{d}^T, \boldsymbol{\gamma}^T|\boldsymbol{y}, \boldsymbol{b}) = \|\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Z}\mathrm{Diag}(\tilde{\boldsymbol{\Gamma}}\boldsymbol{b})(\boldsymbol{1_q} \otimes I_m)\|^2 + \lambda\left(\sum_{j=1}^{p}\frac{|\beta_i|}{|\hat{\beta}_j|} + \sum_{k=1}^{k}\frac{|d_k|}{|\hat{d}_k|}\right), \quad (2.5)$$

where $\hat{\beta}_j$ and $\hat{d}_k$ are ordinary least squares estimates and $\boldsymbol{1_q}$ 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 $\boldsymbol{X_i}$, $\boldsymbol{b_i}$, $\boldsymbol{y_i}$, and $\boldsymbol{\varepsilon_i}$ and setting $\boldsymbol{Z} = \mathrm{diag}(\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_n})$ with corresponding $\tilde{\Pi} = \mathrm{diag}(\Pi, \ldots, \Pi)$, the linear mixed effects model in (2.4) can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\varepsilon}.$$

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

$$Q(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{P_z}(\boldsymbol{y} - \boldsymbol{X\beta}) + n\sum_{j=1}^{p} P_\lambda(|\beta_j|), \quad (2.6)$$

where $\boldsymbol{P_z} = (\boldsymbol{I} + \sigma^{-2}\boldsymbol{Z}\tilde{\Pi}\boldsymbol{Z^T})^{-1}$ and the penalty function $P_\lambda(|\beta_j|)$ is the Smoothed Clipped Absolute Deviation (SCAD, Fan and Li, 2001) penalty with tuning parameter λ. It is important to note that the problem $Q(\boldsymbol{\beta})$ depends on the unknown parameters $\tilde{\Pi}$ and σ^2. To overcome this obstacle, a proxy matrix $\tilde{\boldsymbol{P}}_{\boldsymbol{z}} = (\boldsymbol{I} + \boldsymbol{Z}\,\mathcal{M}\boldsymbol{Z^T})^{-1}$, where $\mathcal{M} = (\log n)\boldsymbol{I}$, is substituted into (2.6) for $\boldsymbol{P_z}$. 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

$$Q(\boldsymbol{b})=\frac{1}{2}(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{b})^T\boldsymbol{P}_{\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{b})+\frac{1}{2}\sigma^2\boldsymbol{b}^{\boldsymbol{T}}\tilde{\boldsymbol{\Pi}}+\boldsymbol{b}+n\sum_{k=1}^{q}P_\lambda(|b_k|), \quad (2.7)$$

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

$$Q(\boldsymbol{b})=\frac{1}{2}(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{b})^T\boldsymbol{P}_{\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{b})+\frac{1}{2}\boldsymbol{b}^{\boldsymbol{T}}\mathscr{M}\boldsymbol{b}+n\sum_{k=1}^{q}P_\lambda(|b_k|).$$

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 $N \le p$, 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

$$\ell_{nM}(\boldsymbol{\beta},\boldsymbol{\Pi},\sigma^2)=-\frac{1}{2}\sum_{i=1}^{n}\log|\sigma^2\boldsymbol{V}_i|-\frac{1}{2}\log\sigma^{-2}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}), \quad (2.8)$$

where $\boldsymbol{V}_i=\boldsymbol{I}_{m_i}+\boldsymbol{Z}_i\boldsymbol{\Pi}\boldsymbol{Z}_i^T$. In high dimensional settings when $N \le p$, the restricted term in (2.8) will become singular so the following full log-likelihood must be used

$$\ell_{nF}(\boldsymbol{\beta},\boldsymbol{\Pi},\sigma^2)=-\frac{1}{2}\sum_{i=1}^{n}\log|\sigma^2\boldsymbol{V}_i|-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}). \quad (2.9)$$

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

$$Q_n(\boldsymbol{\beta},\boldsymbol{\Pi},\sigma^2)=\ell_n(\boldsymbol{\beta},\boldsymbol{\Pi},\sigma^2)-\lambda_1P_1(\beta)-\lambda_2P_2(\boldsymbol{D}), \quad (2.10)$$

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 $\Pi = \boldsymbol{L}\boldsymbol{L}^{\boldsymbol{T}}$ where $\boldsymbol{L}$ is a lower triangular matrix with positive diagonal elements. If any diagonal element of $\boldsymbol{L}_{kk} = 0$, 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

$$Q_n(\boldsymbol{\beta}, \boldsymbol{\Pi}, \sigma^2) = \ell_n(\boldsymbol{\beta}, \boldsymbol{\Pi}, \sigma^2) - \lambda_1 \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_j|} - \sum_{k=2}^{q} \frac{\sqrt{L_{k1}^2 + \ldots L_{kq}^2}}{\|\hat{L}_{(k)}\|}, \quad (2.11)$$

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

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

Chen S, Bowman FD. A Novel Support Vector Classifier for Longitudinal High Dimensional Data. Statistical Analysis and Data Mining. 2011; 4(6):604–611.

Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X, Just M, Newman S. Learning to decode cognitive states from brain images. Machine Learning. 2004; 57:145–175.

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

Mourao-Miranda J, Friston KJ, Brammer M. Dynamic discrimination analysis: a spatial-temporal SVM. Neuroimage. 2007; 36(1):88. [PubMed: 17400479]

Guyon I, Elisseeff A. Introduction to Variable and Feature Selection. Journal of Machine Learning Research. 2003; 3:1157–1182.

Wahba, G. Spline Models for Observational Data. SIAM; Philadelphia, PA: 1990.

Hastie T, Rosset S, Tibshirani R, Zhu J. The Entire Regularization Path for the Support Vector Machine. Journal of Machine Learning Research. 2004; 5:1391–1415.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of American Statistical Association. 2001:1348–1360.

Zou H. The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association. 2006; 101(476):1418–1429.

Bondell H, Krishna A, Ghosh SK. Joint variable selection for fixed and random effects in linear mixed-effects models. Biometrics. 2010; 66:1069–1077. [PubMed: 20163404]

Laird NM, Ware JL. Random-effects models for longitudinal data. Biometrics. 1982; 38:963–974. [PubMed: 7168798]

Laird NM, Lange N, Stram D. Maximum likelihood computations with repeated measures: Application of the EM algorithm. Journal of the American Statistical Association. 1987; 82:97–105.

Fan Y, Li R. Variable selecion in linear mixed effects models. Annals of Statistics. 2012; 40:2043–2068. [PubMed: 24850975]

Efron B, Hastie T, Johnstone I, Tibshirani R. Least Angle Regression. The Annals of Statistics. 2004; 32(2):407–451.

Zou H, Zhang H. On the adaptive elastic-net with a diverging number of parameters. Annals of Statistics. 2009; 37:1733–1751. [PubMed: 20445770]

Li Y, Wang S, Song PX-K, Wang N, Zhu J. Doubly Regularized Estimation and Selection in Linear Mixed-Effects Models fo High-Dimenstional Longitudinal Data. Unpublished manuscript.

Lindstrom MJ, Bates DM. Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. Journal of the American Statistical Association. 1988; 83:1014–1022.

Yuan M, Lin Y. Model Selection and Estimation in Regression With Grouped Variable. Journal of the Royal Statistical Society, Series B. 2006; 68:49–67.

Young DR, Saksvig BI, Wu TT, Zook K, Li X, Champaloux S, Grieser M, Lee S, Treuth M. Multilevel Predictors of Physical Activity For Early, Mid, and Late Adolescent Girls. Journal of Physical Activity & Health. in press.

Bondell H, Reich BJ. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. Biometrics. 2008; 64(1):115–23. [PubMed: 17608783]

Chen Z. Random effects selection in linear mixed models. Biometrics. 2003; 59:762–769. [PubMed: 14969453]

Hall DB, Praestgaard JT. Order-restricted score tests for homogeneity in generalised linear and nonlinear models. Biometrika. 2001; 88(3):739–751.

Lin X. Variance component testing in generlised linear models with random effects. Biometrika. 1997; 84(2):309–326.

Stram DO, Lee JW. Variance components testing in the longitudinal mixed effects model. Biometrics. 1994; 50(4):1171–7. [PubMed: 7786999]

Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B (Methodological). 2005; 67(2):301–320.

Table 1

Comparison of three penalized linear mixed effects models

	1. Joint Penalization	**2. Independent Selection**	**3. Double Penalization**
Authors	Bondell et al. (2010)	Fan and Li (2012)	Li et al. (2012)
Regularizations	1	1	2
Objective Functions	$\ell(\beta, \Pi, \sigma^2) - \lambda P(\beta, \boldsymbol{b})$	Fixed: $\ell(\beta) - \lambda_1 P_1(\beta)$ Random: $\ell(\boldsymbol{b}, \sigma^2)\ \lambda_2 P_2(\boldsymbol{b})$	$\ell(\beta, \Pi, \sigma^2) - \lambda_1 P_1(\beta) - \lambda_2 P_2(\boldsymbol{b})$
Penalties	1 total (ALASSO)	1 fixed (SCAD) 1 random (SCAD)	1 fixed (ALASSO) 1 random (L-2 norm)
Covariance Structure	Modified Cholesky decomposition for covariance matrix	Use of proxy matrix to substitute for unknown covariance matrix	Cholesky decomposition for covariance matrix
Algorithms	EM algorithm	LARS/Elastic net	New efficient algorithm with two quadratic components
High-Dimensional Data	EM algorithm is not efficient for large number of predictors	p and q can diverge to ∞, but must reduce dimensions of fixed effects ignoring random effects first	p and q can diverge to *1*