Published in final edited form as:

Wiley Interdiscip Rev Comput Stat. 2015 ; 7(1): 99–108. doi:10.1002/wics.1322.

Integrative Analysis of “-Omics” Data Using Penalty Functions

Qing Zhao¹, Xingjie Shi^{1,2}, Jian Huang³, Jin Liu⁴, Yang Li⁵, and Shuangge Ma^{1,6,*}

¹Department of Biostatistics, School of Public Health, Yale University

²School of Statistics and Management, Shanghai University of Finance and Economics

³Department of Statistics and Actuarial Science, University of Iowa

⁴Division of Epidemiology and Biostatistics, UIC School of Public Health

⁵School of Statistics, Center for Applied Statistics, Renmin University of China

⁶School of Statistics, Capital University of Economics and Business

Abstract

In the analysis of omics data, integrative analysis provides an effective way of pooling information across multiple datasets or multiple correlated responses, and can be more effective than single-dataset (response) analysis. Multiple families of integrative analysis methods have been proposed in the literature. The current review focuses on the penalization methods. Special attention is paid to sparse meta-analysis methods that pool summary statistics across datasets, and integrative analysis methods that pool raw data across datasets. We discuss their formulation and rationale. Beyond “standard” penalized selection, we also review contrasted penalization and Laplacian penalization which accommodate finer data structures. The computational aspects, including computational algorithms and tuning parameter selection, are examined. This review concludes with possible limitations and extensions.

Keywords

Integrative analysis; omics data; marker selection; penalization

Introduction

Last decade has witnessed an explosion of large-scale datasets, especially in the field of biology and life sciences. The number of features available usually greatly exceeds the number of samples, which has posed a daunting challenge for model estimation and prediction. It has become essential to integrate different sources of data to better separate “signals” from “noises” and the relevant from the irrelevant. Meta-analysis, which in a narrow sense refers to data integration solely based on summary statistics such as effect size, has played an important role in summarizing evidences from for example multiple Genome-Wide Association Studies (Evangelou and Ioannidis (2013); Pharoah et al. (2013); Franke et al. (2010)). Accumulating evidences, in for example Wei et al. (2012); Li et al. (2011); Shen

*Correspondence to: shuangge.ma@yale.edu.

et al. (2013); Sivriver et al. (2011); Zhang et al. (2012); Ma et al. (2011b,a, 2012b), have shown that jointly analyzing multiple datasets can improve analysis results. The scientific findings of these analyses encompass a wide range of topics in the field of biology and biomedicine, including inferring potential biological mechanisms, discovering novel disease subtypes, predicting novel complex biological networks as well as identifying biologically and clinically meaningful cancer prognostic markers, which have opened up new and exciting areas of investigation. The benefits of joint analyses have motivated novel methodological development in the integrative analysis of multiple high-dimensional datasets. Penalized variable selection has been used to synthesize summary statistics that have information on both the estimated effect sizes and covariance matrices. In other scenarios where raw data are available, effective variable selection methods have been employed to achieve a more effective pooling of information.

Penalized variable selection methods, as a family of regularization methods, have been well recognized because of their advantages in reducing noise, improving model prediction, and enhancing model interpretability. In general, regularization refers to the scenario where the estimation for over-parameterized problems can be improved through the use of additional assumptions or prior information. The sparsity condition is widely adopted in “*large p, small n*” problems, which assumes that only a few of the measured features are relevant or carry predictive power. It naturally brings about a penalty function that measures the complexity of a candidate model and encourages sparse structures. Usually a data-dependent tuning parameter leads to the classic tradeoff between overfitting the noise and systematic bias in prediction models. Generically, the penalized estimator can be defined as

$$\hat{\theta} = \operatorname{argmin}_{\theta} \{R(D; \theta) + P(\lambda; \theta)\},$$

where $R(D; \theta)$ is a function of the observed data D and regression coefficients θ and measures the lack-of-fit. It can for example be a negative log-likelihood function or a least squares-type loss function. $P(\lambda; \theta)$ is the penalty function that measures model complexity with tuning parameter λ . Various penalties have been created in single-dataset-analysis settings. The penalization approach faces greater challenges in integrative analysis, where diverse multi-modal datasets come into play. The balance between commonality and data-specificity in model construction has to be coped with fine-tuning in a more subtle way.

This review will focus on diverse penalties developed to accommodate exploitable structures underlying multiple datasets in terms of homogeneity and heterogeneity, applied to integrative analysis with different forms of data. Below we first describe integrative analysis based on summary statistics. Then we discuss the integrative analysis methods when raw data are available. Such methods can conduct one- or bi-level selection as well as have additional smoothing effects. For implementation, we briefly introduce the computational algorithms and tuning parameter selection. The article concludes with discussions and possible extensions.

Data and Model Setting

Integrative analysis can accommodate multiple independent datasets or a single dataset with multiple correlated response variables. Let $M(> 1)$ be the number of datasets (or number of response variables in a single dataset), n^m be the number of subjects in dataset m , and p be the number of covariates. The raw data consist of $\{(y^m, X^m), m = 1, \dots, M\}$, where y^m is the response variable in the m th dataset, and X^m is the corresponding p -vector of covariates. First consider continuous responses. Other types of responses can be analyzed in a similar manner. For dataset m , assume a linear regression model

$$y^m = X^{m'} \beta^m + \varepsilon^m. \quad (1)$$

$X^{m'}$ is the transpose of X^m . $\beta^m = (\beta_1^m, \dots, \beta_p^m)'$ is the regression coefficients for the m th dataset. Denote $\beta_j = (\beta_j^1, \dots, \beta_j^M)$, which represents the effects of j th covariate across M datasets. For the random error ε^m , $E(\varepsilon^m) = 0$, $Var(\varepsilon^m) = \sigma_m^2$. If data are not properly normalized, intercept terms should be added and are not subject to penalization.

Homogeneity and Heterogeneity Models—Under the *homogeneity* model, for any $j (= 1, \dots, p)$, $I(\beta_j^1 = 0) = \dots = I(\beta_j^M = 0)$. That is, if a covariate is identified as associated with response in one dataset, it is identified in all of the M datasets. Thus, the M datasets have the same sparsity structure. One typical example is to integrate data generated in multiple independent studies with comparable designs. For example, multiple datasets have been generated collecting mRNA gene expression measurements and searching for genes associated with the risk of one particular cancer type. On the other hand, the *heterogeneity* model, which includes the homogeneity model as a special case, allows a covariate to be associated with response variables in some datasets, but not others. That is, it is possible that one $I(\beta_j^k = 0) (k = 1, \dots, M)$ is not equal to the remaining. There are multiple scenarios under which the heterogeneity model is meaningful. The first is the analysis of data on different types of diseases. The second is the analysis of data on different subtypes of the same disease. The third scenarios is where subjects in different studies have different environmental exposures or risk factors. The heterogeneity model is more flexible in addressing both common information and data specificity. It is crucial in integrative analysis as it adds one more layer of flexibility and complexity to the existing model construction.

Sparse Meta-Analysis Methods

Sparse meta-analysis is proposed in He (2012) for variable selection in the integrative analysis taking the form of summary statistics. It applies to data integration of, for example, multiple studies with comparable designs. Homogeneity comes from the shared research question and similarity in study designs and data collection, while the sources of heterogeneity include different ways of assessing the omics measurements, different subtypes of diseases, and differences in study population and distributions of effect modifiers across studies. The homogeneity and heterogeneity structures in covariate selection are considered by assuming not only group-level sparsity but also additional

within-group sparsity, which allows each covariate in the important set to be partly active across datasets. Differential penalization is further introduced to incorporate prior knowledge about differential covariate estimates corresponding to various structures. Both of the strategies help extract *homogeneous* information shared by different datasets while allowing for *heterogeneity* among datasets.

1-Norm Group Penalization—In this set of analysis, the only information available pertains to the summary statistics, for example, ordinary least squares estimates $\hat{\beta}^m$ ($m = 1, \dots, M$). It is also assumed that the corresponding variance estimates for individual regression coefficient vectors are available. The goal is to identify correctly the important set in which $\beta^m = 0$ for some m and to estimate the effects of those covariates.

The overall objective function with penalty can be written as

$$(\hat{\beta}^1, \dots, \hat{\beta}^M) = \arg\min_{\beta} \left\{ \sum_{m=1}^M (\hat{\beta}^m - \beta^m)' (\hat{V}^m)^{-1} (\hat{\beta}^m - \beta^m) + \lambda \sum_{j=1}^p \left(\sum_{m=1}^M w_j^m |\beta_j^m| \right)^{1/2} \right\}.$$

λ is a tuning parameter, $\hat{V}^m = \hat{Cov}(\hat{\beta}^m)$ ($m = 1, \dots, M$) is the estimated covariance matrix, and w_j^m is a user-specified weight. In this case, the lack-of-fit measure is the re-scaled squared errors of the regression coefficients. The penalty function achieves additional within-group sparsity by applying the group bridge penalty (Huang et al. (2009)) that is capable of simultaneous group and individual variable selection. Here the group structure is formed by regression coefficients of each covariate across M studies, and the group-level selection determines whether the covariate is associated with the responses at all. For implementation, with similar manipulations mentioned in Huang et al. (2009), the above penalization problem is equivalent to solving

$$\arg\min_{\beta} \left\{ \sum_{m=1}^M (\hat{\beta}^m - \beta^m)' (\hat{V}^m)^{-1} (\hat{\beta}^m - \beta^m) + \sum_{j=1}^p \gamma_j^{-1} \left(\sum_{m=1}^M w_j^m |\beta_j^m| \right) + \lambda_1 \sum_{j=1}^p \gamma_j \right\}, \quad (2)$$

subject to $\gamma_j = 0$ for $j = 1, \dots, p$.

Closely examining the penalties, we can see that the former penalty is an adaptively weighted L_1 penalty. Therefore, the solution is sparse at the within-group individual-covariate level. On the other hand, a small value of γ_j forces the whole group of coefficients to be zero, which leads to group selection. The 1-norm group penalties are flexible in that they free the covariates from being completely active across all the M datasets, compared to the group Lasso method employed in single-dataset settings.

Differential Penalization—Differential penalization using penalty weights has become widely noticed through adaptive Lasso. The adaptive Lasso provides an effective remedy for the inconsistency of Lasso selection which forces the coefficients to be equally penalized. Adaptive weights are used to penalize different coefficients differently, and the adaptive

Lasso has been shown to enjoy the oracle properties under some data and model settings. In general, the flexibility in pre-specifying differential penalization provides a flexible way to incorporate prior knowledge about different structures among datasets.

Data-dependent penalty weights are adopted as

$$w_j^m = |\hat{\beta}_j^m|^{-1}. \quad (3)$$

It has been pointed out that these weights are critical to consistency of the method. The intuition is that as sample size grows, the weights for zero-coefficient predictors get inflated, whereas those for nonzero-coefficient predictors converge to finite constants. Thus asymptotically unbiased estimates for large coefficients and thresholding estimates for small coefficients can be achieved simultaneously. The penalty can also be modified to encompass the homogenous assumption. Under this scenario, all the M coefficients for covariate j are

subject to a common weight, denoted as $w_j = (\sum_{m=1}^M |\hat{\beta}_j^m|/M)^{-1}$. It is noted that this assumption is different from the homogeneity model we assume previously as it cannot achieve common sparsity structure (across datasets) as both group-level and individual-level selection has been conducted.

There are situations where differential penalties can be used to describe the subgroup structures among datasets. Say that there are H subgroups in the M datasets, each studying a subtype of the disease. Denote S_h as the set of datasets in the h th subgroup. The objective function is the same as above except that the penalty weights are chosen as

$$w_j^m = \left\{ \sum_{h=1}^H [I(m \in S_h) |S_h|^{-1} \sum_{l \in S_h} |\hat{\beta}_j^l|] \right\}^{-1}, j=1, \dots, p; m=1, \dots, M, \quad (4)$$

where $|S_h|$ is the cardinality of S_h . It has been shown that this approach has a better chance of identifying the true model and better performance in estimation and prediction.

Penalized Integrative Analysis Methods

With improving technologies and communication, it has become more feasible to obtain raw data from individual studies. At the same time, an increasing number of networks or consortia have been created to share the original data, which largely accelerates the advancement of integrative analysis. By pooling information from multiple raw datasets, penalized integrative analysis methods have been shown to outperform metaanalysis with more accurate marker identification and better prediction. The asymptotic equivalence in relative efficiency of using summary statistics versus individual-level data is established only under a fix-effect model that assumes the effect sizes are the same for all studies. Penalized marker selection methods have been extensively studied and are flexible in coping with the complex homogeneous and heterogeneous structures both in covariate selection and effect size estimation. The homogeneity and heterogeneity models in marker selection are accommodated with similar rationale but different penalty functions. Extra smoothing

effects over the regression coefficients can be pursued to encourage similarity in effect sizes both across and within datasets.

As above, the overall objective function consists of two parts, the lack-of-fit measure and the penalty function. Still consider continuous responses and linear regression models. Denote y as the length $\sum_m n^m$ vector composed of all responses, X as the $(\sum_m n^m) \times (Mp)$ design matrix (for all datasets and all subjects), and β as the Mp vector composed of β^m 's. Consider the overall lack-of-fit measure

$$(y - X\beta)' \Omega (y - X\beta). \quad (5)$$

Ω is a symmetric matrix that does not depend on β . The basic requirement for Ω is that under the “classic” asymptotics with $\sum_m n^m \gg p$, minimizing the above lack-of-fit measure with respect of β leads to a consistent estimate. Ω is especially useful in the analysis of data with multiple correlated responses. Since this review mostly focuses on the penalty functions as opposed to loss functions, we assume the simplified condition with $\Omega = I$.

Composite Penalization—Composite penalization refers to using penalties that possess a composite of two-layer of penalties that are decoupled with each other, with the outer penalty determining how groups are related to each other and the inner penalty dictating the relationship within groups. Penalties that fall into this category include the composite absolute penalty family that is proposed in Zhao et al. (2009) for grouped and hierarchical variable selection. This family utilizes the distinct behaviors of the bridge estimates, more specifically according to different γ -norm ($\gamma = 1$ or > 1) penalized estimates, to describe the across-group and within-group structures. We refer to the original publication for more details. The aforementioned 1-norm group penalization can be thought of as a composite penalty with a group-level bridge penalty and an individual-level L_1 penalty. This penalty framework has made it possible to perform bi-level selection that accommodates both *homogeneity* and *heterogeneity* in a natural and flexible manner.

Heterogeneity Model: Under the heterogeneity model, first consider the composite MCP (Minimax Concave Penalty), which is shown in Breheny and Huang (2009) to have the oracle properties and satisfactory performance in single-dataset analysis. MCP is defined as

$\rho(t; \lambda, \gamma) = \lambda \int_0^{|t|} \left(1 - \frac{x}{\lambda\gamma}\right)_+ dx$. $\lambda > 0$ is the tuning parameter, as with many other penalties. $\gamma > 0$ is the regularization parameter. The composite MCP penalty for integrative analysis proposed in Liu et al. (2014) can be expressed as

$$P_{cMCP}(\beta) = \sum_{j=1}^p \rho_O \left(\sum_{m=1}^M \rho_I(|\beta_j^m|; \lambda, a); \sqrt{M_j} \lambda, b \right). \quad (6)$$

Both ρ_O and ρ_I are MCP. λ denotes the tuning parameter, and a, b denote possibly different regularization parameters. M_j is the number of datasets that measure covariate j and accommodates partially matched covariate sets. The two-layer structure is more obvious

when expressed in the partial derivative with respect to β_j^m which is

$\rho'_o \left(\sum_{m=1}^M \rho_I(|\beta_j^m|; \lambda, a); \sqrt{M_j} \lambda, b \right) \rho'_I(|\beta_j^m|; \lambda, a)$. The outer penalty ρ_O applies to a sum of inner penalties ρ_I , which are formed by the M coefficients of covariates. Though this method shares similar rationale with the group bridge method, it has been noted that it may achieve consistency in the selection of individual variables because of the advantage of MCP over Lasso (Huang et al. (2012)).

Homogeneity Model: The homogeneity model emphasizes the commonality among different datasets. If there is strong evidence supporting the assumption, the homogeneity model can be preferred over the heterogeneity model owing to its simplicity in modeling and computation. The composite penalty decouples the bi-level selection so that it can be easily adapted to the homogeneity model. The 2-norm group MCP penalty has been proposed as

$$P_{gMCP}(\beta) = \sum_{j=1}^p \rho(\|\beta_j\|_2; \sqrt{M_j} \lambda, \gamma). \quad (7)$$

Notations have similar definitions as above. We note that the outer penalty remains to be MCP, but the inner penalty is modified to be the L_2 norm (square root of the ridge penalty) which only has the shrinkage but not selection property. This penalty achieves the one-layer selection that is enough to determine whether a covariate is associated with the responses at all.

Contrasted Penalization—Contrasted penalization imposes penalties on the differences among a defined coefficient profiles, thus forcing a certain level of similarity among different effect sizes. The smoothing concept can be traced back to the term “fusion”, borrowed from Land and Friedman (1996). It is first used to exploit the spatial nature of predictor variables in signal and image processing and has become widely known through the fused Lasso. With fused Lasso, the penalty is written in the form of $\lambda \sum_j |\beta_j - \beta_{j-1}|$, the sum of L_1 differences between adjacent coefficients. It attaches after a sparsity-encouraging Lasso penalty and produces a simple piecewise constant solution. For computational simplicity, the existing contrasted penalization methods have chosen a quadratic form of smoothing. When concatenating with the aforementioned penalties, the contrasted penalization methods belong to a large family of penalties that uses a combination of L_1 - and L_2 -type penalties, in a similar spirit as the Smooth-Lasso (S-Lasso) type methods defined in Hebiri and van de Geer (2011). When the contrast penalty shrinks the differences between the effect sizes of the same covariate in multiple datasets, it encourages a finer structure for *homogeneity across datasets* and improves prediction performance.

In Shi et al. (2014), the following contrast penalty is imposed in addition to bi-level group selection penalties (for example, the group bridge penalty)

$$\lambda \sum_{j=1}^p \sum_{(k,l): k \neq l} a_{jl}^{kl} (\beta_j^k - \beta_j^l)^2. \quad (8)$$

$\lambda = 0$ is a data-dependent tuning parameter, and $a_{jl}^{kl} = I\{\text{sgn}(\beta_j^k) = \text{sgn}(\beta_j^l)\}$. Here $\text{sgn}(\cdot)$ is the sign function. When $\text{sgn}(\beta_j^k) \neq \text{sgn}(\beta_j^l)$ (one positive and another negative, or one nonzero and another zero), covariate j demonstrates different effects in different datasets. On the other hand, when $\text{sgn}(\beta_j^k) = \text{sgn}(\beta_j^l)$, particularly when both of them are nonzero, covariate j has qualitatively similar effects in dataset k and l , and the smoothing effect allows them to borrow information from each other. There are many ways of extracting prior knowledge about the signs. One approach is to conduct penalized estimation for each dataset and determine the signs based on whether the coefficient is zero or nonzero and whether the signs match if both of them are nonzero. Another approach is to estimate the marginal effect of each covariate for each dataset. The asymptotic consistency of these approaches has been discussed. In practice, the penalized estimation approach is shown to have better performance.

Laplacian Penalization—The grouping effect among highly correlated covariates poses significant challenges to selection, especially under the $p \gg n$ setting. This problem has been addressed in single-data analysis by methods such as Enet and Mnet in Huang et al. (2010) (which is a combination of MCP and L_2 penalties). However, when the dimensionality is high and data structures are complex, as often taking place in the integrative analysis of multiple datasets, it is with a higher possibility that spurious correlations lead to false positive signals (Fan et al. (2013)). An effective way of dealing with the collinearity among predictors is to incorporate specific prior information on the correlation patterns. Laplacian penalization adopts a quadratic penalty that allows augmented knowledge of the data structures and encourages smoothness among the coefficients associated with correlated predictors. With an extra level of smoothness in effect size among predictors, *homogenous signals within the same datasets* can be largely enhanced. The Laplacian penalization also belongs to the family of S-Lasso type estimations but outperforms by not incurring extra bias.

Formulated in a network analysis paradigm, Liu et al. (2013a) imposes the following Laplacian penalty in addition to a sparse group penalty

$$\frac{1}{2} \tau \times M \times \sum_{1 \leq k < l \leq p} a_{kl} \left(\frac{\|\beta_k\|_2}{\sqrt{M_k}} - \frac{\|\beta_l\|_2}{\sqrt{M_l}} \right)^2, \quad (9)$$

where $\tau = 0$ is a data-dependent tuning parameter. Denote

$\theta = (\theta_1, \dots, \theta_p)' = \left(\frac{\|\beta_1\|_2}{\sqrt{M_1}}, \dots, \frac{\|\beta_p\|_2}{\sqrt{M_p}} \right)'$. The above penalty can be expressed using a positive semi-definite matrix L , which satisfies $\theta' L \theta = \sum_{1 \leq k < l \leq p} a_{kl} (\theta_k - \theta_l)^2$. Let $A = (a_{kl}, 1 \leq k, l \leq p)$ and $G = \text{diag}(g_1, \dots, g_p)$, where $g_k = \sum_{l=1}^p a_{kl}$. Under a network analysis framework where

a_{kl} is the weight of edge (k, l) , g_j is the degree of vertex j . The penalty above can then be written in a “network friendly” manner as $\frac{1}{2}\tau M\theta' L\theta$ where $L = G - A$. Here the Laplacian matrix is not normalized, meaning that the weight g_j is not standardized to 1. In problems where predictors should be treated without preference with respect to network connectivity, the matrix can be normalized such that $L^\star = I_p - A^\star$ with $A^\star = G^{-1/2}AG^{-1/2}$ and the penalty can be modified to $\theta L^\star \theta$. For discussions on normalization specifically, refer to Liu et al. (2013a).

There are multiple ways of constructing the adjacency matrix A , which measures the strength of connection between any two nodes and is the most important characteristic of a network. Denote a_{kl} as each of the elements. Denote r_{kl} as the absolute value of a simple correlation coefficient (Pearson, rank, etc). Many commonly used adjacency measures take the following forms: (a) $a_{kl} = I(r_{kl} > r)$, where r is a data-dependent cutoff; (b) $a_{kl} = r_{kl}^\alpha$ with $\alpha > 0$ and can be determined using for example the scale-free topology criterion (Zhang and Horvath (2005)); and (c) $a_{kl} = r_{kl}^\alpha I(r_{kl} > r)$. We refer to Horvath (2011); Liu et al. (2013a) for more discussions. Measures in the form of (a) and (c) are sparse in that some components may be exactly zero while measures in the form of (b) are “soft” adjacency measures. With the adjacency matrix, the interplay among different covariates can be effectively accommodated into integrative analysis, and the Laplacian penalization approach has been shown to have superior performance over the alternatives.

Computation

Computational Algorithms—The above penalization methods can be realized using group coordinate descent (GCD) algorithms first proposed by Yuan and Lin (2006), which are built on coordinate descent (CD) algorithms. CD algorithms are first proposed by Fu (1998); Friedman et al. (2007); Wu and Lange (2008) for optimization problems with convex penalties and later extended to optimization with non-convex penalties. A CD algorithm optimizes an objective function with respect to a single parameter at a time, and iteratively cycles through all parameters until convergence is reached. GCD algorithms treat a single group at a time instead of a single parameter. A generic GCD algorithm for integrative analysis can be described in the following steps:

Step 1 Initialize $s = 0$, the estimate $\beta^{(0)} = (0, \dots, 0)'$, and the vector of residuals $r = y - X\beta^{(0)}$.

Step 2 For $j = 1, \dots, p$, consider updating with the j th group

- a. Calculate $z_j = n^{-1} X_j' r + \beta_j^{(s)}$. X_j is the submatrix of X that corresponds to β_j .
- b. Update $\beta_j^{(s+1)} \leftarrow F(z_j; \lambda)$. $F(z_j; \lambda)$ denotes the soft-threshold operator used to obtain the minimizers, depending on the penalty applied. λ represents the tuning parameter.
- c. Update $r \leftarrow r - X_j(\beta_j^{(s+1)} - \beta_j^{(s)})$.

Step 3 Update $s \leftarrow s + 1$.

Step 4 Repeat Step 2-3 until convergence.

This framework can be ready to compute various penalized estimates by adjusting Step 2 (a) and (b). For example, with the penalty which is a combination of group MCP and Laplacian penalty under the homogeneity model as in Liu et al. (2013a), consider the overall objective function

$$L(\beta; \lambda, \gamma) = \frac{1}{2n} \|\mathbf{y} - \sum_{j=1}^p \mathbf{X}_j \beta_j\|^2 + \sum_{j=1}^p \rho(\|\beta_j\|_2; \sqrt{M_j} \lambda_1, \gamma) + \frac{1}{2} \lambda_2 M \sum_{1 \leq k < l \leq p} a_{kl} \left(\frac{\|\beta_k\|_2}{\sqrt{M_k}} - \frac{\|\beta_l\|_2}{\sqrt{M_l}} \right)^2. \quad (10)$$

For $j = 1, \dots, p$, given the group parameters $\beta_k (k = j)$ fixed, to minimize the objective function $L(\beta; \lambda, \gamma)$ with respect to the j th group parameter β_j is equivalent to minimizing

$$R(\beta_j) = \frac{1}{2} a \beta_j' \beta_j - \mathbf{b}' \beta_j + c \|\beta_j\| + C, \quad (11)$$

where C is a term free of β_j , and (i) when $\|\beta_j\| \leq \gamma \lambda_1 \sqrt{M_j}$, $a = 1 + \lambda_2 \sum_{l: l \neq k} \frac{a_{kl}}{M_l} - \frac{1}{\gamma}$,

$\mathbf{b} = \frac{1}{n} \mathbf{X}_j' \mathbf{r} + \beta_j^{(s)}$, and $c = \lambda_1 \sqrt{M_j} - \frac{\lambda_2 M}{\sqrt{M_j}} \sum_{l: l \neq k} \frac{a_{kl}}{M_l} \|\beta_k\|$; and (ii) when $\|\beta_j\| > \gamma \lambda_1 \sqrt{M_j}$,

$a = 1 + \lambda_2 \sum_{l: l \neq k} \frac{a_{kl}}{M_l}$, $c = -\frac{\lambda_2 M}{\sqrt{M_j}} \sum_{l: l \neq k} \frac{a_{kl}}{M_l} \|\beta_k\|$, and $\mathbf{b}$ is the same as under the previous situation. Then the minimizer of $R(\beta_j)$ is

$$\beta_j = \frac{1}{a} \left(1 - \frac{c}{\|\mathbf{b}\|} \right)_+ \mathbf{b}. \quad (12)$$

The modification is as follows:

Step 2

a. Calculate $a, \mathbf{b}, c$ as in expression (11).

b. Update $\beta_j^{(s+1)}$ using expression (12).

In terms of the convergence properties, the objective function can be rewritten in the form of

$f(\beta) = f_0(\beta) + \sum_{j=1}^p f_j(\beta_j)$, where

$$f_0(\beta) = \frac{1}{2n} \|\mathbf{y} - \sum_{j=1}^p \mathbf{X}_j \beta_j\|^2 + \frac{1}{2} \lambda_2 M \sum_{1 \leq k < l \leq p} a_{kl} \left(\frac{\|\beta_k\|_2}{\sqrt{M_k}} - \frac{\|\beta_l\|_2}{\sqrt{M_l}} \right)^2;$$

$f_j(\beta_j) = \rho(\|\beta_j\|_2; \sqrt{M_j} \lambda_1, \gamma)$. Since f_0 is regular in the sense of Tseng (2001) and

$\sum_{j=1}^p f_j(\beta_j)$ is group-wise separable, according to Theorem 4.1(c) of Tseng (2001), the GCD solution converges to a coordinate-wise minimum point of f , which is also a stationary point.

In the case of composite MCP penalization under the heterogeneity model as in Liu et al. (2014), in Step 2 (b), a local coordinate descent approach (LCD), as proposed in Breheny and Huang (2011), is adopted in addition to an inner cycling procedure of each of the M parameters within the j th group, as composite MCP does not have a convenient form of updating individual groups. Step 2 (b) is modified to

Step 2 (b) For $m = 1, \dots, M$, update $\beta_j^{m(s+1)} \leftarrow f_{cMCP}(z_j^m; \lambda_{jm})$, where z_j^m is the m th element of z_j .

Here $f_{cMCP}(z_j^m; \lambda_{jm}) = S_1(z_j^m; \lambda_{jm})$, where $S_1(z; \lambda) = \text{sgn}(z)(|z| - \lambda)_+$, and

$\lambda_{jm} = \rho' \left(\sum_{m=1}^M \rho(|\beta_j^m|; \lambda_2, b); \lambda_1, a \right) \rho'(|\beta_j^m|; \lambda_2, b)$. λ_{jm} is obtained by taking the first order Taylor series approximation of β_j so that the penalty as a function of β_j^m is approximately proportional to $\lambda_{jm} |\beta_j^m|$. It is noted that λ_{jm} is updated in each iteration with the most current estimate. Refer to Liu et al. (2014) for more details on the convergence properties of the algorithms.

Tuning and Regularization Parameter Selection—The tuning parameter λ controls the trade-off between model fitting and sparsity and is crucial to performance of the methods. The tuning parameter is selected from a sequence of values. The solutions over a grid of values between λ_{min} and λ_{max} are computed, with λ_{max} estimated through making $\beta = 0$ solution for the objective function. Then different model selection criteria can be applied to choose an optimal regularized model. Tibshirani (1996) first proposed the generalized cross validation (GCV) type criterion

$$GCV(\hat{f}) = \frac{1}{n} \sum_{i=1}^n \left[\frac{y_i - \hat{f}(x_i)}{1 - p(t)/n} \right]^2.$$

$\hat{f}(x_i)$ is an estimate of y_i . $p(t)$ is the effective number of parameters and can be approximated by $p(t) = \text{tr}\{X(X^T X + \lambda W)^{-1} X^T\}$ ($W = \text{diag}(|\beta_j|)$). However, it has been noticed that the asymptotic behavior of GCV is very similar to the traditional model selection criterion AIC (Wang et al. (2007)), which is not an efficient and consistent selection criterion and leads to potential overfitting. Yet it is well known that BIC is consistent under “standard” conditions. The following extended BIC (Chen and Chen (2008)) introduces an additional term considering both the number of parameters and complexity of model space to restrict the model selection behavior of BIC:

$$BIC_\gamma(s) = n \log \hat{\sigma}_s^2 + \nu(s) \log n - 2\gamma \log \{p(s)\}.$$

$\hat{\sigma}_s^2$ is the estimated variance of random errors, $\nu(s)$ is the number of components in model s , and $p(s)$ is the prior probability of model s , which is proportional to the inverse of the size of model s defined. Modified BIC-type tuning parameter selectors, which involve the estimates

of effective number of parameters, have been pursued by Wang et al. (2007) for non-convex penalized regression under the assumption of fixed predictor dimension and by Wang et al. (2009) with a diverging number of parameters. We refer to Zou et al. (2007) and Fan and Li (2001) for detailed discussions on estimating degrees of freedom for regularization methods. The tuning parameter selection criterion used in sparse meta-analysis is a modified BIC criterion. Good estimates for the effective number of parameters in variable selection are still very challenging.

Another popular method for selecting tuning parameter is K -fold cross-validation. Most of the currently available R packages for penalized methods provide solution paths for different tunings with K -fold cross-validation, as only simple calculations not overall iterations are needed and cross-validation is affordable. However, overfitting has been noted in cross-validation in high-dimensional settings. Feng and Yu (2013) propose a new procedure called consistent cross-validation (CCV). CCV is shown to enjoy the consistency properties under certain conditions.

Furthermore, for some penalties, we encounter the selection of regularization parameters besides tuning parameters. For composite MCP, in (6), smaller values of a and b are better at retaining the unbiasedness of MCP penalties for large coefficients, but they also have the risk of creating objective functions with a nonconvexity problem that are difficult to optimize and yield solutions that are discontinuous with respect to λ . It is therefore advisable to choose values of a and b that are big enough to avoid this problem but not too big. As suggested in Breheny and Huang (2011) and Zhang (2010), we recommend in numerical studies experimenting with a few values for a and b , particularly including 1.8, 3, 6, and 10.

Conclusions

In this article, we have focused on integrative analysis based on summary statistics and raw data from multiple datasets using penalty functions. We have reviewed various penalization methods, their rationales, and their roles in integrative analysis, and summarized the algorithms and tuning parameter selection in realizing them. Integrative analysis has posed a great challenge to model estimation with the complex and exploitable structures of multiple datasets. The homogeneity and heterogeneity in covariate selection can be accommodated through differential 1-norm group penalties and one-dimensional and two-dimensional composite penalization respectively in a flexible way. The homogeneity in across-dataset and within-dataset covariate estimates can be further enhanced through smoothing effects of contrasted penalization and Laplacian penalization. Prior knowledge is incorporated naturally to separate “signals” from “noises”. Brief comparisons between the methods have been discussed. We have also reviewed the GCD computational algorithms, which belong to a large family, to implement the methods. The convergence properties of most of the algorithms can be derived using Tseng (2001). Composite MCP and some other penalties require extra approximations of the penalties to obtain a locally manageable representation. We refer to Mazumder et al. (2011); Tseng and Yun (2009); Friedman et al. (2007) and others for non-smooth and non-separable coordinate descent algorithms. Selecting tuning parameters using different criteria has been reviewed. Recently, Meinshausen and Bühlmann (2010) proposes stability selection, which is based on resampling. It would be interesting to

apply it and compare with the existing methods. In addition, there are several recent studies on the identification of gene-environment interactions using penalization (Liu et al. (2013c); Shi et al. (2014)). Penalization methods achieve simultaneous estimation and identification of gene-environment interactions without resorting to significance levels, which are subject to multiple-comparison adjustment in high-dimensional settings. With flexible penalties in exploiting grouping structures, Liu et al. (2013c) effectively describes the hierarchical relationship between main effects and interactions by grouping one SNP (as a functional unit) and its interactions with environmental risk factors. We recognize that it is still challenging to accommodate with affordable computational cost the prohibitively large number of higher-order interaction terms especially gene-gene interactions in joint modeling. There is much left open to further research in interaction modeling in integrative analysis.

For presentation simplicity, we have limited our discussions of penalty functions to the scenario with multiple independent datasets. The penalties are relatively “independent” of the loss functions. The aforementioned methods are applicable to the analysis of multiple correlated outcomes with modifications of the loss functions. For variable selection methods as the main focus, we refer to Liu et al. (2013b). For dimension reduction and estimation, we refer to Yuan et al. (2007) and others. In integrative analysis, there are other groups of researchers carrying out un-supervised clustering analysis. Shen et al. (2013) use latent variable regression methods for joint modeling of multiple omics data types including mRNA gene expression and copy number variation to identify common latent variables that can be used to cluster patients into biologically and clinically relevant disease subtypes. Different penalty functions are incorporated additionally to produce multiple denoised versions of centroids. Sivriver et al. (2011) model time-course gene expression data with mixture priors and conduct clustering with penalization techniques adopted to reduce overfitting. We point out that there is a lack of semi-supervised and supervised statistical learning methods integrating multiple omics data or time-course gene expression data. We foresee that flexible and complex penalty functions can play a promising role in future research on these topics. Due to space limitation, we leave them and others (Costa et al. (2007); Gross and Tibshirani (2014)) for further reading. Also, current integrative analysis mainly assumes parallel structures among the multiple related data information. There is a need for methodological development in penalized integrative analysis to accommodate hierarchical structures among different layers of information. Another limitation is that the theoretical aspect has not been well developed. For a few specific models and penalization methods (Liu et al. (2013b); Ma et al. (2012a)), we have shown that they enjoy the selection and estimation consistency properties. Especially this result holds if $\log(p)/N \rightarrow 0$ as $N, p \rightarrow \infty$, where N is the combined sample size across datasets. More theoretical investigations are needed. Last but not least, we barely touch upon topics about statistical inference in high-dimensional settings post model selection. We refer to Zhang and Zhang (2014); Berk et al. (2013); Javanmard and Montanari (2014); Taylor et al. (2014); van de Geer et al. (2013); Leeb et al. (2014); Voorman et al. (2014) for the most recent research on significance tests and confidence intervals of regression coefficients among others.

Acknowledgments

We thank the editors, associate editor, and reviewers for invitation, careful review and insightful comments. Our research has been supported by NIH awards CA142774, CA165923, and CA182984, National Social Science Foundation of China (13CTJ001), and National Natural Science Foundation of China (71301162).

References

- Berk R, Brown L, Buja A, Zhang K, Zhao L. Valid post-selection inference. *The Annals of Statistics*. 2013; 41(2):802–837.
- Breheny P, Huang J. Penalized methods for bi-level variable selection. *Stat Interface*. 2009; 2(3):369–380. [PubMed: 20640242]
- Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*. 2011; 5(1):232–253. [PubMed: 22081779]
- Chen J, Chen Z. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*. 2008; 95(3):759–771.
- Costa I, Krause R, Opitz L, Schliep A. Semi-supervised learning for the identification of syn-expressed genes from fused microarray and in situ image data. *BMC Bioinformatics*. 2007; 8(Suppl 10):S3. [PubMed: 18269697]
- Evangelou E, Ioannidis JPA. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet*. 2013; 14(6):379–389. [PubMed: 23657481]
- Fan, J.; Han, F.; Liu, H. Challenges of big data analysis. 2013. <http://arxiv.org/abs/1308.1479>
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96(456):1348–1360.
- Feng, Y.; Yu, y. Consistent cross-validation for tuning parameter selection in high-dimensional variable selection. 2013. <http://arxiv.org/abs/1308.5390>
- Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, Anderson CA, Bis JC, Bumpstead S, Ellinghaus D, Festen EM, Georges M, Green T, Haritunians T. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nat Genet*. 2010; 42(12):1118–1125. [PubMed: 21102463]
- Friedman J, Hastie T, Höfling H, Tibshirani R. Pathwise coordinate optimization. *The Annals of Applied Statistics*. 2007; 1(2):302–332.
- Fu WJ. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*. 1998; 7(3):397–416.
- Gross, SM.; Tibshirani, R. Collaborative regression. 2014. <http://arxiv.org/abs/1401.5823>
- He, Q. PhD thesis. University of North Carolina; Chapel Hill: 2012. Variable Selection, Sparse Meta-analysis and Genetic Risk Prediction for Genome-wide Association Studies.
- Hebiri M, van de Geer S. The smooth-lasso and other l1+l2 -penalized methods. *Electronic Journal of Statistics*. 2011; 5:1184–1226.
- Horvath, S. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer; 2011.
- Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models. *Statistical Science*. 2012; 27(4):481–499.
- Huang, J.; Breheny, P.; Ma, S.; Zhang, CH. Technical Report 402. Department of Statistics, University of Iowa; 2010. The mnet method for variable selection. <http://www.stat.uiowa.edu/files/stat/techrep/tr402.pdf>
- Huang J, Ma S, Xie H, Zhang CH. A group bridge approach for variable selection. *Biometrika*. 2009; 96(2):339–355. [PubMed: 20037673]
- Javanmard, A.; Montanari, A. Confidence intervals and hypothesis testing for high-dimensional regression. 2014. <http://arxiv.org/abs/1306.3171>

- Land, SR.; Friedman, JH. Technical Report 656. Department of Statistics, Stanford University; 1996. Variable fusion: A new adaptive signal regression method. <http://www.stat.cmu.edu/tr/tr656/tr656.html>
- Leeb, H.; Poetscher, BM.; Ewald, K. On various confidence intervals post-model-selection. 2014. <http://arxiv.org/abs/1401.2267>
- Li W, Liu CC, Zhang T, Li H, Waterman MS, Zhou XJ. Integrative analysis of many weighted co-expression networks using tensor computation. *PLoS Comput Biol*. 2011; 7(6):e1001106. [PubMed: 21698123]
- Liu J, Huang J, Ma S. Incorporating network structure in integrative analysis of cancer prognosis data. *Genetic Epidemiology*. 2013a; 37(2):173–183. [PubMed: 23161517]
- Liu J, Huang J, Zhang Y, Lan Q, Nothman N, Zheng T, Ma S. Integrative analysis of prognosis data on multiple cancer subtypes using penalization. *Biometrics*. 2013b In press.
- Liu J, Huang J, Zhang Y, Lan Q, Rothman N, Zheng T, Ma S. Identification of gene–environment interactions in cancer studies using penalization. *Genomics*. 2013c; 102(4):189–194. [PubMed: 23994599]
- Liu J, Ma S, Huang J. Integrative analysis of cancer diagnosis studies with composite penalization. *Scandinavian Journal of Statistics*. 2014; 41(1):87–103.
- Ma S, Dai Y, Huang J, Xie Y. Identification of breast cancer prognosis markers via integrative analysis. *Computational Statistics and Data Analysis*. 2012a; 56(9):2718–2728. [PubMed: 22773869]
- Ma S, Huang J, Song X. Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics*. 2011a; 12(4):763–775. [PubMed: 21415015]
- Ma S, Huang J, Wei F, Xie Y, Fang K. Integrative analysis of multiple cancer prognosis studies with gene expression measurements. *Statistics in Medicine*. 2011b; 30(28):3361–3371. [PubMed: 22105693]
- Ma S, Zhang Y, Huang J, Huang Y, Lan Q, Rothman N, Zheng T. Integrative analysis of cancer prognosis data with multiple subtypes using regularized gradient descent. *Genetic Epidemiology*. 2012b; 36(8):829–838. [PubMed: 22851516]
- Mazumder R, Friedman JH, Hastie T. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*. 2011; 106(495):1125–1138. [PubMed: 25580042]
- Meinshausen N, Bühlmann P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2010; 72(4):417–473.
- Pharoah PDP, Tsai YY, Ramus SJ, Phelan CM, Goode EL, Lawrenson K, Buckley M, Fridley BL, Tyrer JP, Shen H, Weber R, Karevan R, Larson MC. Gwas meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nat Genet*. 2013; 45(4):362–370. [PubMed: 23535730]
- Shen R, Wang S, Mo Q. Sparse integrative clustering of multiple omics data sets. *The Annals of Applied Statistics*. 2013; 7(1):269–294. [PubMed: 24587839]
- Shi X, Liu J, Huang J, Zhou Y, Xie Y, Ma S. A penalized robust method for identifying gene–environment interactions. *Genetic Epidemiology*. 2014; 38(3):220–230. [PubMed: 24616063]
- Sivriver J, Habib N, Friedman N. An integrative clustering and modeling algorithm for dynamical gene expression data. *Bioinformatics*. 2011; 27(13):392–400.
- Taylor, J.; Lockhart, R.; Tibshirani, RJ.; Tibshirani, R. Post-selection adaptive inference for least angle regression and the lasso. 2014. <http://arxiv.org/abs/1401.3889>
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*. 1996; 58:267–288.
- Tseng P. Convergence of a block coordinate descent method for non-differentiable minimization. *Journal of Optimization Theory and Applications*. 2001; 109(3):475–494. Communicated by O.L. Mangasarian.
- Tseng P, Yun S. A coordinate gradient descent method for non-smooth separable minimization. *Mathematical Programming*. 2009; 117(1-2):387–423.
- van de Geer, S.; Bühlmann, P.; Ritov, Y.; Dezeure, R. On asymptotically optimal confidence regions and tests for high-dimensional models. 2013. <http://arxiv.org/abs/1303.0518>

- Voorman, A.; Shojaie, A.; Witten, D. Inference in high dimensions with the penalized score test. 2014. <http://arxiv.org/abs/1401.2678>
- Wang H, Li B, Leng C. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2009; 71(3):671–683.
- Wang H, Li R, Tsai CL. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*. 2007; 94:553–568. [PubMed: 19343105]
- Wei Y, Li X, Wang Qf, Ji H. iaseq: integrative analysis of allele-specificity of protein-dna interactions in multiple chip-seq datasets. *BMC Genomics*. 2012; 13(1):681. [PubMed: 23194258]
- Wu TT, Lange K. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*. 2008; 2(1):224–244.
- Yuan M, Ekici A, Lu Z, Monteiro R. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2007; 69(3):329–346.
- Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2006; 68(1):49–67.
- Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*. 2005; 4:45.
- Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*. 2010; 38(2):894–942.
- Zhang CH, Zhang SS. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2014; 76(1): 217–242.
- Zhang S, Liu CC, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research*. 2012; 40(19):9379–9391. [PubMed: 22879375]
- Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*. 2009; 37:3468–3497.
- Zou H, Hastie T, Tibshirani R. On the “degrees of freedom” of the lasso. *The Annals of Statistics*. 2007; 35(5):2173–2192.