Multivariate Regression with Small Samples: A Comparison of Estimation Methods

W. Holmes Finch **Maria E. Hernández Finch**
Ball State University

High dimensional multivariate data, where the number of variables approaches or exceeds the sample size, is an increasingly common occurrence for social scientists. Several tools exist for dealing with such data in the context of univariate regression, including regularization methods (i.e., Lasso, Elastic net, Ridge Regression, as well as Bayesian models with spike and slab priors. These methods have not been widely studied in the context of multivariate regression modeling. Thus, the goal of this simulation study was to compare the performance of these methods for high dimensional data with multivariate regression, in which there exist more than one dependent variable. Simulation results revealed that the regularization methods, particularly Ridge Regression, were found to be particularly effective in terms of parameter estimation accuracy and control over the Type I error rate. Implications for practice are discussed.

Social scientists frequently work in contexts with multiple dependent variables of interest, where appropriate data analysis involves the use of multivariate linear models. In some situations, the number of independent variables (*p*) may approach, or even exceed the sample size (*N*), leading to what is commonly referred to as high dimensional data. When used with high dimensional data, standard regression estimators, including those associated with multivariate models, yield unstable coefficient estimates with inflated standard errors (Bühlmann & van de Geer, 2011), leading to reduced statistical power and erroneous conclusions regarding relationships between independent and dependent variables. Furthermore, when *p* exceeds *N*, it is simply not possible to obtain estimates for model parameters using standard estimation methods. The problems associated with high dimensional data in the univariate case could be further amplified when the data are multivariate in nature, given that the number of parameters to be estimated is the number of independent variables +1 multiplied by the number of dependent variables. Although prior research has been done focusing on methods for dealing with high dimensional univariate linear models, relatively little work has been done in the context of multivariate linear models. Therefore, the objective of this simulation study was to compare the performance of several methods for handling high dimensional multivariate data with one another, and with standard ordinary least squares (OLS) multivariate regression. First, a description of OLS regression is provided, followed by descriptions of models designed for use in the high dimensional case, including the lasso, elastic net, and ridge regression. Next, descriptions of two Bayesian alternatives for multivariate regression estimation are provided. The research goals and the methodology used to address those goals are then presented, followed by a discussion of the results of the simulation study, and an application of each method to an existing dataset. Finally, the implications of the simulation results, in light of existing research, are discussed.

Ordinary Least Squares Regression

The multivariate linear regression model can be written as:

$$\boldsymbol{y}_i = \beta_0 + \beta_{1p}x_{1i} + \beta_{2p}x_{2i} + \cdots + \beta_{jp}x_{ji} \tag{1}$$

where $\boldsymbol{y}_i$ = Vector of dependent variables for subject *i*
x_{ji} = Independent variable *j* for subject *i*
β_0 = Intercept
β_{jp} = Coefficient for independent variable *j* on dependent variable *p*

To obtain estimates for the model coefficients ($\hat{\beta}$), the least squares (LS) estimator is typically used. LS identifies $\hat{\beta}$ values that minimize the squared residuals of the model in (1), as expressed in equation (2).

$$e^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2}$$

where *N*=Total sample size
$\hat{\boldsymbol{y}}_i = \hat{\beta}_0 - \hat{\beta}_{1p}x_{1ip} + \hat{\beta}_{2p}x_{2ip} + \cdots + \hat{\beta}_{jp}x_{jip}$
$\hat{\beta}_0$ = Estimate of model intercept
$\hat{\beta}_{jp}$ = Estimate of coefficient for independent variable *j* on dependent variable *p*

Regularization Methods

As noted above, the presence of high dimensional data can result in estimation problems for the OLS estimator, rendering it less than optimal in such cases (Bühlmann & van de Geer, 2011). There exist several alternatives for use when dealing with high dimensional data in the context of linear models, including variable selection methods (e.g., stepwise regression, best subsets regression), and data reduction techniques (e.g., principal components regression). Research has found that with high dimensional data, variable selection methods produce inflated standard errors for model coefficients (Hastie, Tibshirani, & Friedman, 2009). Data reduction techniques mitigate this problem by combining independent variables into a small number of linear combinations, but make interpretation of results for individual variables difficult (Finch, Hernandez Finch, & Moss, 2014).

A third family of approaches for regression with high dimensional data involves parameter estimation algorithms known as regularization, or shrinkage techniques. Variable selection methods assign inclusion weights of either 1 (include variable in model) or 0 (exclude variable from model) to each independent variable, and then estimate $\hat{\beta}_{jp}$ for each included variable. Regularization methods identify optimal values of $\hat{\beta}_{jp}$ such that the most important independent variables receive higher values, and the least important are assigned coefficients at or near 0. Researchers have found that the resulting standard errors do not suffer from the inflation inherent with variable selection methods (Hastie et al., 2009). Additionally, regularization methods avoid the increased complexity associated with data reduction techniques, by not merging the individual independent variables into linear combinations. These regularization methods that have been shown to be effective for univariate regression (Zou & Hastie, 2005; Tibshirani, 1996). However, prior research has not examined their performance in the context of multivariate regression, which is the focus of the current study.

Lasso

Regularization methods work by applying a penalty to the OLS estimator from equation (1). One such approach, the least absolute shrinkage and selection operator (lasso; Tibshirani, 1996) is expressed as:

$$e^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \lfloor \hat{\beta}_j \rfloor \tag{3}$$

The terms in equation (3) are as defined in (2), with the addition of the parameter λ, which controls the degree to which the model coefficients are down weighted or removed from the model (shrinkage). Larger λ values correspond to greater shrinkage, and when λ=0, the lasso is simply the OLS estimator. Lasso is designed to eliminate from the model independent variables that contribute very little to the explanation of the dependent variable, by setting their $\hat{\beta}$ values to 0, while at the same time retaining independent variables that are important in explaining y. The optimal value of λ is identified using jackknife cross-validation, which is described in Tibshirani.

Elastic Net

A second regularization method that can be used with high dimensional data is the elastic net, (EN; Zou & Hastie, 2005), which expands upon the Lasso penalty function by including a second shrinkage parameter, α, as part of the fitting function. EN minimizes equation (4):

$$e_i^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p}\left(\alpha \lfloor \beta_j \rfloor + (1-\alpha)\beta_j^2\right) \tag{4}$$

The value of α is selected using cross-validation techniques in the same manner as for λ.

Ridge Regression

A third regularization method that is closely associated with both lasso and EN is ridge regression (RR; Hoerl, 1962). Indeed, RR is a close variant of EN, such that when α = 0 in equation (4) the EN model simplifies to the RR model. Thus, the RR penalty minimizes the function:

$$e_i^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \beta_j^2 \tag{5}$$

Bayesian Regression

Each of the methods described above relies on the frequentist approach to parameter estimation. An alternative set of methodologies rests on Bayesian estimation, in which prior information about the distributions of the model parameters is combined with information from the data to create a posterior distribution for each parameter; e.g., regression coefficient. These estimates are obtained using the

Markov chain Monte Carlo (MCMC) approach (see Kaplan, 2014 for a detailed description of MCMC). MCMC samples a very large number (e.g., 10,000) of random draws from the posterior distribution for each parameter in the model, which is itself constructed using a markov chain. McNeish (2016), among others, has commented on the advantages inherent within the Bayesian framework when researchers are working with small sample sizes, and complex model structures, such as those that are present with high dimensional data. By incorporating prior information about the distributions of the model parameters, estimation in such difficult situations may be possible for Bayes where it is not for the frequentist models (McNeish). However, it is also key that in such situations, appropriate prior distributions be put on the parameters, otherwise McNeish points out that Bayesian estimates might actually be less accurate and less efficient than those produced in the frequentist context.

When considering the prior distributions to be used, the researcher has two choices: informative and noninformative priors. In the context of linear regression, common noninformative prior distributions for the regression coefficients are the uniform $(-\infty, \infty)$ or the normal with a mean of 0 and variance of 1000 (Kaplan, 2014; Kruschke, 2015). Such noninformative priors allow for the least impact of the prior on the posterior distributions of the parameters. In contrast, perhaps the most common informative prior distribution is the normal with a given mean and a small variance, such as 0.1. The mean of the informative prior would be based upon previous results that are known to the researcher, such as through publication in the literature. Of course, frequently in practice we do not have sufficient information to warrant using such informative priors, and thus rely instead on the noninformative variety. This will be the case in the current study. Finally, the point estimate for a given parameter is typically taken as either the mean or median of the posterior distribution (Kaplan).

Bayesian Regression with Spike and Slab Priors

As discussed briefly above, Bayesian estimation presents an alternative approach for fitting regression models. Within this framework, a Bayesian approach that has been suggested for use with high dimensional data structures involves use of the spike and slab prior distribution for model parameter estimates (Ishwaran & Rao, 2005). Rockova and George (2016) described what they termed the spike and slab lasso (SS), which builds upon prior work in the area of Bayesian variable selection (Kyung, Gill, Ghosh, & Casella, 2010). SS is very similar to the Bayesian methodology described above, with the difference lying in the nature of the prior distribution used. Rather than basing the priors on the normal or uniform distributions, SS uses a combined set of priors known collectively as the spike and slab. The key quality of the SS approach is the use of two prior distributions for each model coefficient. The first of these is the spike, which essentially identifies irrelevant model coefficients; i.e. those that are likely to be 0 in the population. Thus, in the end, each model parameter will be assigned either a 1 (relevant) or 0 (irrelevant) value in the posterior of the spike distribution. The slab distribution corresponds only to those effects that are deemed relevant by the spike. Therefore, the slab prior distribution serves much the same role as the prior in standard Bayesian regression, as described previously. And indeed, the slab distribution will generally employ mean and variance values that are commonly seen in the context of standard Bayesian regression.

The spike and slab algorithm operates as a two stage process. First, the posterior of the spike distribution is obtained for each coefficient. Those coefficients with a posterior centered on 1 are considered to be relevant, and are thus retained into the second step of the model, where they are assigned a normal prior with a given mean (often 0) and variance. Combined, the SS prior has the following form:

$$p(\beta_j|r_j) = (1 - r_j)\delta_0 + r_j(N(0, \sigma^2)) \tag{6}$$

where, $p(r_j) = Bernoulli(p_j)$; i.e., hyperparameter for the probability that coefficient j is relevant δ_0 =Point mass function at 0. Equation (6) illustrates that when the distribution of the spike is at or very close to 0 (i.e., $r_j \rightarrow 0$), the $p(\beta_j|r_j)$ will be 0. On the other hand, when $r_j \rightarrow 1$, $p(\beta_j|r_j)$ will be estimated as the posterior based upon the data and the normally distributed prior, much as was the case with the standard Bayesian regression model described previously.

Previous research has shown that SS yields univariate regression model parameter estimates with relatively low bias, and more accurate standard errors than the standard lasso, under many conditions (Xu & Ghosh, 2015). Similar results for factor loadings were also reported by Lu, Show, and Loken (2016).

Although promising in the context of univariate linear models, the performance of SS has not heretofore been explored in the context of multivariate regression. Therefore, it was included in the current study.

Study Goals

The primary goal of this study was to compare the performance of several methods for estimating multivariate regression models in the context of high dimensional data (small samples with multiple independent and dependent variables). Specific outcome variables to be examined were convergence rates, absolute parameter estimation bias, standard errors for the estimates, Type I error rates, and power. In addition, each method was also applied to an existing high dimensional dataset, in order to demonstrate their utility in an applied context.

Methodology

The study goals outlined above were addressed using a Monte Carlo simulation study with 1000 replications per combination of manipulated study conditions, which are described below. Data were generated using a multivariate linear regression model, as in equation (1), with dependent and independent variables both being generated from the $N(0,1)$ distribution. Values of β_j were set to 1 (for non-zero coefficients) or 0. As an example, for the 6 independent variables, 66% non-zero coefficient condition, the data generating model was:

$$\boldsymbol{y}_i = 1 + 1x_{1i} + 1x_{1i} + 1x_{1i} + 1x_{1i} + 0x_{1i} + 0x_{1i} \quad (7)$$

All data were generated in the R software environment, version 3.2.2 (R Foundation for Statistical Computing, 2015). The following variables were manipulated in the study, and were selected to reflect a variety of conditions likely to be seen in practice.

Manipulated variables

- Number of dependent variables: 2, 4, 6
- Number of independent variables: 3, 6, 18
- Sample size: 10, 20, 30, 50, 100
- Correlation among dependent variables: 0, 0.2, 0.5, 0.8
- Percent of non-zero coefficients: 33%, 66%, 100%.
- Estimation methods:
  - OLS
  - Lasso
  - EN
  - RR
  - Bayes with noninformative normal priors (NB)
  - Bayes with spike and slab priors (SS).

The OLS models were fit using the R function `lm`, the lasso, EN, and RR models were fit using the `glmnet` function in the `glmnet` R library. Bayesian regression with normal priors was fit using the `rmultireg` function in the `bayesm` R library, and Bayesian regression with the spike and slab priors was fit using the `MBGLSS` function in the `MBSGS` R library. With regard to the lasso, EN, and RR, the optimal settings for $\lambda$and α were identified through minimization of the mean squared error using 10-fold cross validation, as recommended in Hastie et al. (2009). For both Bayesian estimators, a total of 20,000 draws were made from the Markov chain, with the first 5,000 serving as the burn in interval. The chains were thinned such that every 10th draw was sampled, creating a total of 15,000 data points for parameter estimation. The medians of these posterior distributions were used to obtain the point estimates for each parameter. Preliminary analyses with each of the simulation conditions revealed that these settings uniformly resulted in proper convergence for each of the parameter estimates. For the normal Bayes estimator, a noninformative prior distribution of $N(0, 1000)$ was used. Noninformative priors were also used for the SS estimator, with $p_j = 0.5$ for the spike, and $N(0, 1000)$ used for the slab.

Outcome Variables

The outcomes of interest were convergence rates, absolute parameter estimation bias, standard error of the estimate, and the Type I error and power rates. Convergence rates were simply the proportion of simulation replications for which each estimate converged on a solution. Parameter estimation bias was calculated as:

$$Bias = \left|\beta - \hat{\beta}\right| \tag{8}$$

where: β =Data generating coefficient value and $\hat{\beta}$ =Estimated coefficient value.

Bias was calculated for each coefficient for each of the estimators. The standard error was the standard deviation of the parameter estimates taken across replications, and again was calculated for each coefficient for each of the estimators. The final two outcome variables of interest in this study were the Type I error and power rates for the coefficients. In this study, Type I error refers to the case where a coefficient was found to be significantly different from 0 for a simulation replication, when the data generating value is 0 in the population. Similarly, power was the proportion of cases in which a method identified a parameter as being statistically different from 0 (i.e., significant) when the population generating value was not 0.

In order to identify significant main effects and interactions of the manipulated factors with respect to each of the outcomes, analysis of variance (ANOVA) was used, and both statistical significance and effect sizes in the form of partial omega-squared (ω^2) values were reported. Effects that were both statistically significant and with ω^2 values in excess of 0.1 were identified as being substantively important. The threshold of 0.1 for ω^2 was selected because it corresponds to a model effect accounting for at least 10% of the variance in an outcome variable.

Applied Example

In order to demonstrate the utility of the regularization approaches for fitting linear models with real data, analysis was conducted using an exemplar dataset. The data were collected on 10 adults with autism who were clients of an autism research and service provision center at a large Midwestern university. Adults identified with autism represent a particularly difficult population from which to sample, meaning that quite frequently sample sizes are small. The sample for this analysis was comprised of 10 adults (9 males), with a mean age of 20 years, 2 months (SD=1 year, 9.6 months). Of interest for the current analysis was the relationship between executive functioning as measured by the Delis-Kaplan Executive Functioning System (DKEFS; Delis, Kaplan, & Kramer, 2001) and cognitive ability scores, based on the Wechsler Adult Intelligence Scale, 4th edition (WAIS-IV; Wechsler, 2008). Specifically, scores on the WAIS-IV verbal comprehension, perceptual reasoning, working memory, and processing speed composite scores served as the dependent variables in this multivariate regression analysis. Because of the difficulty in obtaining samples of adults with autism, relatively little work has been conducted with this population regarding the relationship between executive functioning and IQ, although it is known to be particularly relevant for individuals with autism in general (Mclean, Johnson, Zimak, Joseph, & Morrow, 2014). In order to demonstrate the various models featured in this research, the 4 WAIS-IV composite scores were treated as the dependent variables, and the 16 DKEFS subscales appearing in Table 3 as the independent variables. Each estimation method was then fit to the data using the settings described above for the simulation portion of the study.

Results

Convergence Rate

With respect to the convergence rates, the lasso, EN, RR, NB, and SS had 100% convergence rates across all simulated conditions. OLS could not converge for any of the 18 independent variables and sample size of 10 conditions. In addition, when there were 18 independent variables and 6 dependent variables, OLS converged only 25% of the time for *N*=20, 48% of the time for *N*=30, and 100% of the time for *N* of 50 or 100. With 18 independent variables and 4 dependent variables, this convergence rate improved to 67%, 98% for samples of 20, and 30, respectively. When there were 18 independent and 2 dependent variables, OLS converged 100% of the time across conditions, other than for *N*=10. When convergence was not attained, additional replications of the simulations were conducted until the desired 1000 converged solutions were obtained.

Absolute Parameter Estimation Bias

ANOVA identified the interaction of the number of dependent variables by the number of independent variables by the sample size by the estimation method as the highest order term that was statistically significantly related to absolute parameter estimation bias $\left(F_{80,1890} = 10.29, p < 0.001, \omega^2 = 0.303\right)$. All other model terms were either not significant with $\omega^2 > 0.1$, or were subsumed in one of these interactions. Table 1 contains the absolute parameter estimation bias by the number of dependent variables, number of independent variables, sample size, and estimation method. Perhaps most notable among these results is the degree of bias present for OLS when there were 18 independent variables and the sample size was 20. Bias was greater with more dependent variables. The other methods also displayed greater estimation bias when there were 18 independent variables and samples of 10 or 20, but the level of such bias was much lower than that exhibited by OLS. Among the alternative methods, the least bias was in evidence for the lasso, EN, and RR estimators followed by SS, with NB exhibiting more bias than all estimators except for OLS, in this combination of conditions. However, when the sample size was 30 or more, in conjunction with 18 independent variables, estimation bias was lowest for OLS and NB in the 2 and 4 dependent variables conditions. Finally, when there were 18 independent variables, SS exhibited greater bias than the regularization methods for samples of 10, 20, and 30, but had comparable values, and in some cases even less bias, for samples of 50 and 100.

When there were 2 or 6 independent variables, OLS and NB had the least biased coefficient estimates, across sample sizes. The lasso, EN, and RR estimators all had similar levels of bias, which were higher in these conditions than for OLS and NB. SS had greater bias than any of the methods for samples of 30 or fewer, but for samples of 50 or 100 estimation bias for SS was lower than was the case for the lasso, EN, or RR techniques. Finally, for 6 dependent variables and 2 or 6 independent variables, OLS and NB exhibited the lowest estimation bias for all sample sizes, SS had the greatest bias for 6 independent variables with 6 dependent variables, and for 2 independent variables with N of 30 or less. However, for *N* of 100 SS yielded less biased estimates than the lasso, EN, or RR estimators.

Table 1. Absolute Parameter Estimation Bias by Number of Dependent Variables, Independent Variables, Sample Size, and Estimation Method

dv	iv	*N*	OLS	Lasso	EN	Ridge	NB	SS
2	2	10	.0155	.1642	.1735	.1725	.0156	.2646
		20	.0099	.0852	.0936	.1004	.0124	.1002
		30	.0151	.0657	.0711	.0794	.0110	.0519
		50	.0005	.0533	.0583	.0650	.0002	.0335
		100	.0056	.0348	.0407	.0469	.0022	.0137
	6	10	.0244	.2589	.2638	.2672	.0292	.3651
		20	.0022	.1056	.1187	.1244	.0031	.1638
		30	.0028	.0787	.0861	.0925	.0032	.0787
		50	.0047	.0610	.0672	.0737	.0050	.0416
		100	.0038	.0323	.0379	.0437	.0037	.0113
	18	10	NA	.6751	.6279	.6924	.7684	.7337
		20	1.1310	.2743	.2929	.2905	.4210	.3312
		30	.0169	.1100	.1195	.1278	.0184	.2232
		50	.0173	.0476	.0583	.0673	.0119	.0561
		100	.0040	.0363	.0427	.0492	.0018	.0162

Table 1 (continued). Absolute Parameter Estimation Bias by Number of Dependent Variables, Independent Variables, Sample Size, and Estimation Method

dv	iv	*N*	OLS	Lasso	EN	Ridge	NB	SS
4	2	10	.0194	.2225	.2391	.2879	.0215	.3866
		20	.0017	.1716	.1327	.1721	.0010	.1699
		30	.0036	.0813	.1032	.0899	.0040	.0950
		50	.0068	.0632	.0797	.0513	.0069	.0397
		100	.0036	.0409	.0585	.0234	.0037	.0244
	6	10	.0084	.7685	.2444	.2666	.0111	.4045
		20	.0121	.1891	.1527	.1753	.0127	.2267
		30	.0011	.0777	.1055	.1279	.0007	.1405
		50	.0017	.0595	.0813	.1001	.0020	.0622
		100	.0028	.0348	.0513	.0650	.0028	.0215
	18	10	NA	.7683	.7803	.7951	.9195	.8175
		20	2.0713	.2683	.2803	.2951	.5195	.3306
		30	.0023	.0956	.1238	.1505	.0013	.2440
		50	.0033	.0575	.0828	.1045	.0030	.0904
		100	.0079	.0476	.0644	.0799	.0080	.0423
6	2	10	.0037	.2097	.2631	.2955	.0055	.4909
		20	.0049	.1120	.1554	.1927	.0056	.2838
		30	.0033	.0794	.1197	.1498	.0028	.1617
		50	.0031	.0560	.0870	.1138	.0029	.0671
		100	.0005	.0384	.0604	.0810	.0007	.0315
	6	10	.0119	.2632	.2722	.3015	.0054	.6722
		20	.0106	.0983	.1488	.1859	.0099	.6115
		30	.0056	.0875	.1299	.1610	.0060	.5964
		50	.0018	.0639	.0961	.1257	.0020	.4144
		100	.0024	.0421	.0664	.0875	.0026	.3562
	18	10	NA	.9585	.9181	.9273	1.4771	1.1810
		20	3.4226	.2754	.2894	.3137	.6601	.3656
		30	.5097	.1007	.1471	.1817	.3289	.2307
		50	.5132	.0539	.0919	.1228	.3288	.0783
		100	.5117	.0369	.0620	.0839	.3345	.0223

Note. NA=Model could not provide estimates.

Standard Error

The ANOVA for the standard error of the estimates identified the interaction of the proportion of non-zero parameters in the population by estimation method $\left(F_{20,1508} = 19.723, p < 0.001, \omega^2 = 0.207\right)$, and the interaction of number of independent variables by number of dependent variables by method $\left(F_{80,1890} = 8.909, p < 0.001, \omega^2 = 0.279\right)$ as statistically significant, with an effect size greater than 0.1. All other terms were either not statistically significant, had effect size values less than 0.1, or were subsumed within one of these two interactions.

Figure 1 contains the standard error by estimation method and proportion of non-zero coefficients. From these results, it is clear that the lasso yielded the smallest standard error across conditions, whereas both OLS and NB had the largest (and comparable to one another) standard errors. The standard errors for EN, RR, and SS increased with a larger proportion of non-zero population coefficient values, with those of SS being larger than those of the other regularization methods across conditions, and comparable to those of OLS and NB for 66% and 100% non-zero coefficients.

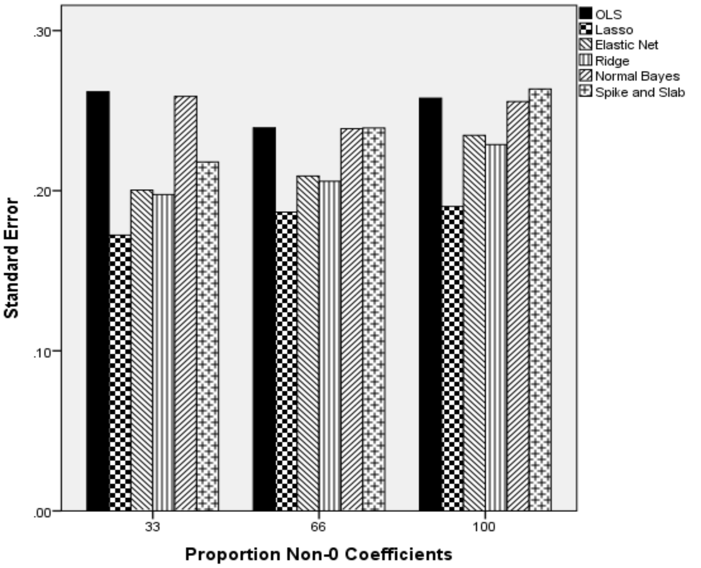

Figure 1. Standard error of regression coefficients by proportion of non-zero coefficients in the model, and estimation method.

Table 2 contains the parameter standard errors by number of dependent and independent variables, the sample size, and the estimation method. When the sample size was 10 or 20 and there were 3 or 6 independent variables in the model, OLS, NB, and SS yielded larger standard error estimates than did the other methods. This pattern was more marked for larger numbers of independent and dependent variables. When there were 18 independent variables, SS produced smaller standard errors than did OLS or NB. As was noted earlier, OLS could not yield estimates when there were 18 independent variables and a sample size of 10. Among the regularization methods, lasso, EN, and RR provided comparable standard error values across conditions, and lower than those of the other methods with samples of 10 or 20. The SS standard errors were comparable to those of the regularization methods when the sample size was 30 or more. Finally, the standard errors for all of the methods were comparable when the sample size was 30 or greater, except for the 18 independent variables. In that case, the OLS and NB standard errors were larger than those of the other methods unless the sample size was at least 50, or in the case of 6 dependent variables, at least 100.

Table 2. Parameter Estimation Bias by Number of Dependent Variables, Independent Variables, Sample Size, and Estimation Method

dv	iv	*N*	OLS	Lasso	EN	Ridge	NB	SS
2	3	10	.4385	.3947	.3981	.3962	.4366	.4556
		20	.2547	.2558	.2464	.2528	.2543	.2663
		30	.1952	.1965	.1908	.1940	.1950	.1908
		50	.1457	.1457	.1429	.1459	.1456	.1637
		100	.1020	.1028	.1020	.1034	.1020	.1013
	6	10	.7273	.4071	.4041	.4055	.6984	.4648
		20	.2979	.2835	.2736	.2646	.2976	.3354
		30	.2096	.2024	.1969	.1943	.2096	.2003
		50	.1538	.1519	.1487	.1463	.1538	.1479
		100	.1025	.1035	.1012	.1003	.1024	.1009
	18	10	NA	.4200	.4280	.4102	.8832	.4689
		20	2.0727	.2692	.2692	.2872	.9952	.3009
		30	.3174	.2030	.2080	.2066	.3164	.1899
		50	.1767	.1650	.1635	.1645	.1765	.1675
		100	.1112	.1094	.1080	.1071	.1112	.1070

Table 2 (continued). Parameter Estimation Bias by Number of Dependent Variables, Independent Variables, Sample Size, and Estimation Method

dv	iv	*N*	OLS	Lasso	EN	Ridge	NB	SS
4	3	10	.4483	.3796	.3637	.3764	.4423	.4430
		20	.2496	.2442	.2282	.2478	.2721	.3227
		30	.1928	.1865	.1787	.1887	.2019	.1590
		50	.1469	.1451	.1408	.1464	.1477	.1027
		100	.1011	.1040	.0984	.1048	.1046	.0736
	6	10	.9543	.3783	.3734	.3501	.9326	.4489
		20	.2908	.2439	.2553	.2418	.2903	.3350
		30	.2068	.1997	.1921	.1891	.2066	.1942
		50	.1508	.1471	.1430	.1401	.1507	.1228
		100	.1053	.1038	.1016	.1009	.1052	.1044
	18	10	NA	.4056	.4169	.3890	1.6039	.4408
		20	2.5654	.2264	.2169	.2290	.9511	.4232
		30	.3055	.2743	.2562	.2438	.3044	.3425
		50	.1814	.1679	.1637	.1612	.1813	.2934
		100	.1093	.1066	.1044	.1037	.1093	.1068
6	3	10	.4590	.3811	.3629	.3387	.4566	.4334
		20	.2601	.2861	.2915	.2903	.2598	.3386
		30	.1936	.1928	.1796	.1754	.1935	.1946
		50	.1523	.1443	.1450	.1409	.1522	.1204
		100	.0998	.1001	.0988	.0970	.0997	.0803
	6	10	1.5312	.4033	.3813	.3900	1.5838	.3533
		20	.3700	.2459	.2307	.2201	.3699	.2389
		30	.2098	.1987	.1895	.1842	.2096	.1659
		50	.1566	.1486	.1445	.1409	.1565	.1260
		100	.1021	.1002	.0976	.0966	.1021	.1010
	18	10	NA	.4361	.4409	.4136	1.7470	.4494
		20	2.9031	.3870	.3927	.3579	2.8968	.4249
		30	.3402	.2756	.2568	.2459	.3200	.2637
		50	.2369	.1685	.1631	.1600	.1808	.1717
		100	.0823	.1052	.1035	.1026	.1085	.1084

Note. NA=Model could not provide estimates.

Type I Error Rate

ANOVA identified the interaction of number of independent variables by sample size by estimation method $(F_{40,1890} = 17.835, p < 0.001, \omega^2 = 0.274)$ to be significantly related to the Type I error rate for the coefficients. Figure 2 displays the Type I error rate by number of independent variables (each in a separate panel) and sample size by method. Per recommendations by Bradley (1978), Type I error rates between 0.025 and 0.075 were considered to be within control. Across number of independent variables, the Type I error rates for OLS and NB were elevated above the nominal 0.05 level, and above 0.075 for

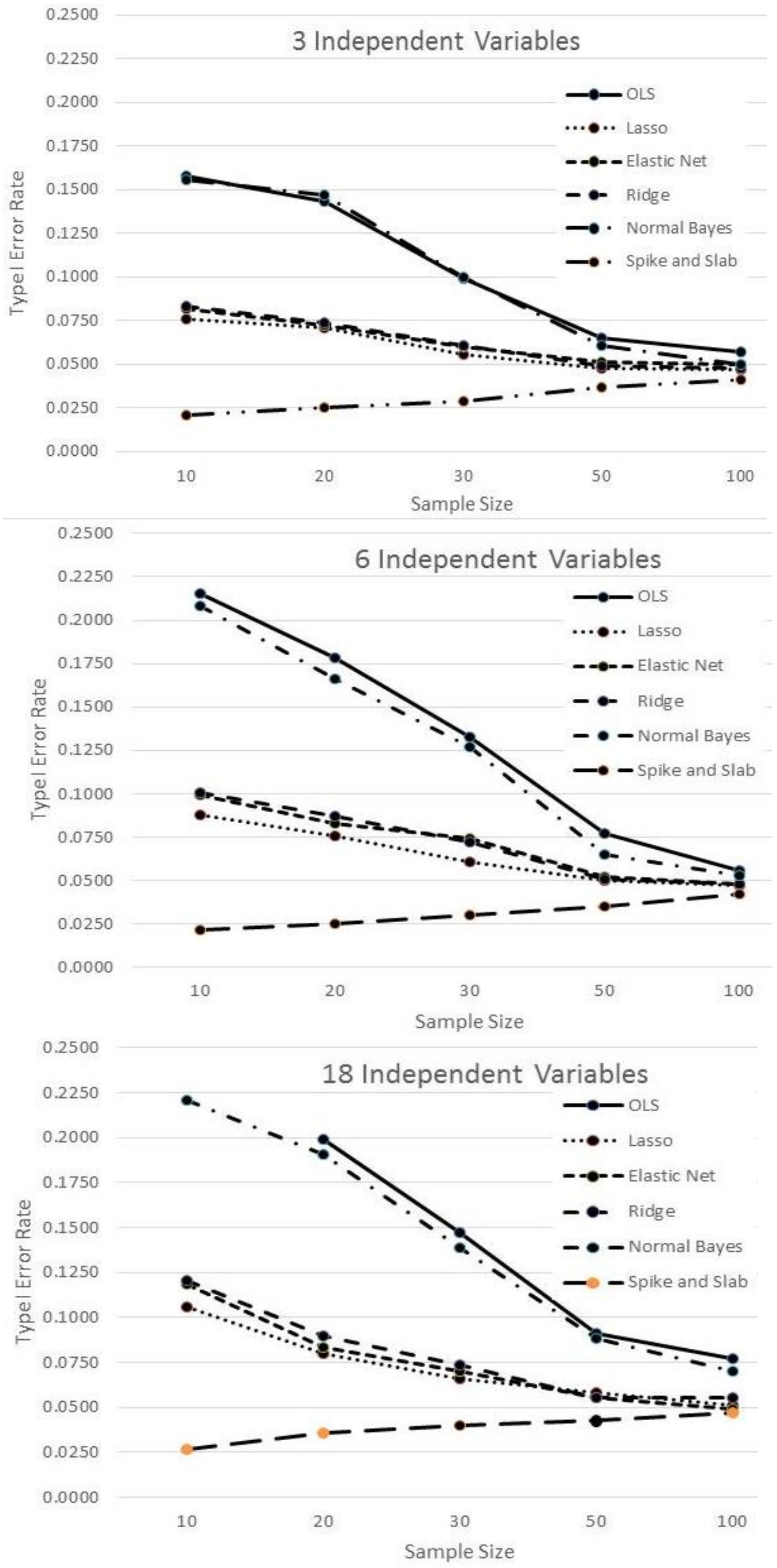

Figure 2. Type I error rate by sample size and number of independent variables.

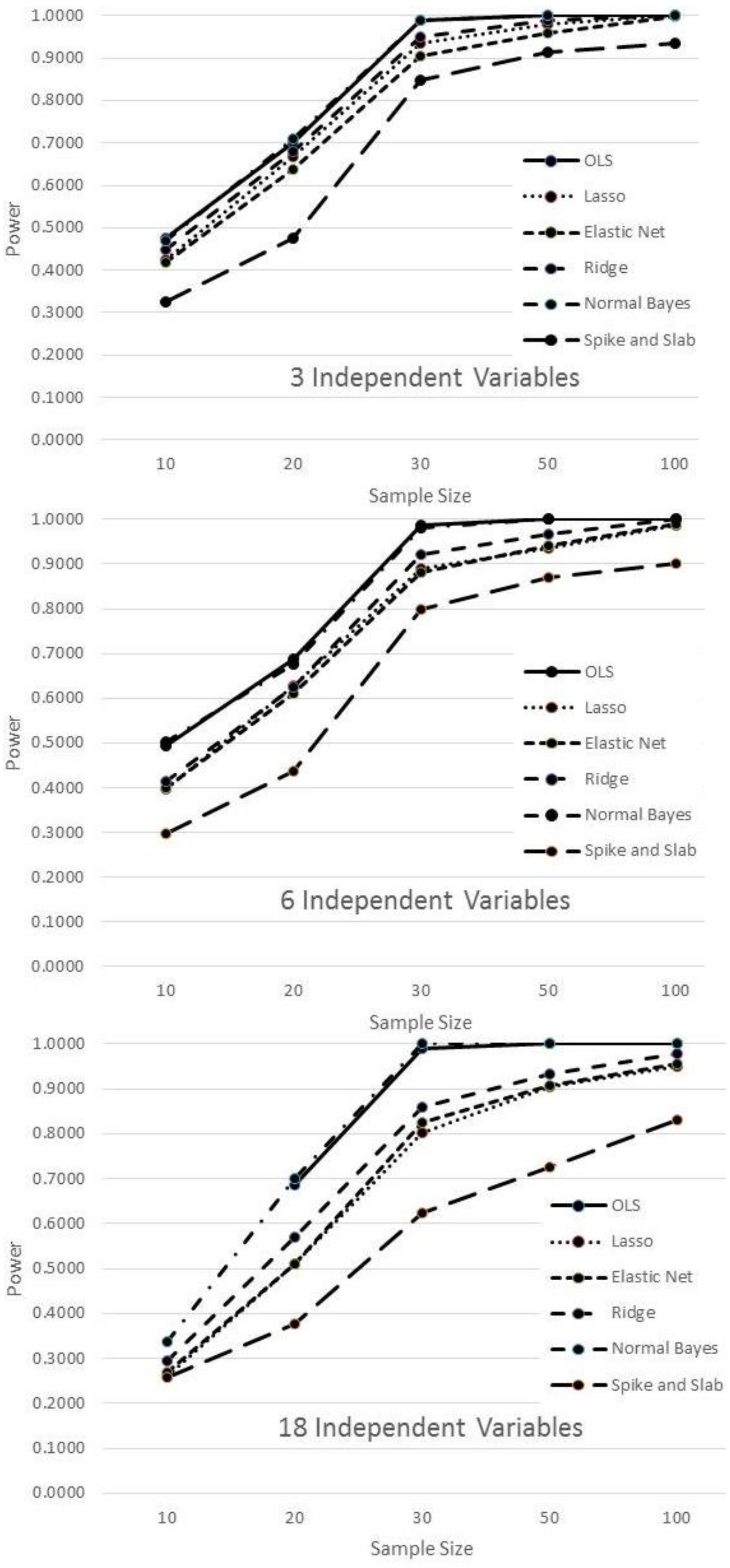

Figure 3. Power by Sample Size and Number of Independent Variables.

samples of fewer than 50. The error rate declined concomitantly with increases in sample size. When there were 18 independent variables, OLS had a Type I error rate greater than 0.05 regardless of sample size, and NB only controlled error for a sample of 100. In contrast, the Type I error rate for SS was always below the nominal 0.05 level, and increased slightly with increases in sample size. The error rates for the lasso, EN, and RR estimators were always lower than those of OLS and NB. In addition, based on the guidelines in Bradley (1978), they were appropriately controlled across sample sizes for 3 independent variables. For 6 independent variables, RR maintained control over the Type I error rate for all sample size conditions except 10, whereas the lasso and EN had elevated rates for samples of 10 and 20, but maintained control over the error rate for *N* of 30 or more. Finally, for 18 independent variables none of the methods except for SS maintained control of the Type I error rate for samples of 10 or 20, but for 30 or more the lasso, EN, and RR all maintained control over the error rate with 18 independent variables. In all cases, the regularization methods yielded lower Type I error rates than did OLS or NB.

Power

Power rates by sample size, number of independent variables, and estimation method appear in Figure 3. Across all number of variables conditions, SS had lower power than did the other methods studied here. This difference in power between SS and the other techniques was greatest at smaller sample sizes, but persisted even for *N* as large as 100. On the other hand, across most conditions displayed in Figure 3 OLS and NB displayed the highest power values. It is important to remember, however, that the Type I error rates for these two methods exceeded the nominal 0.05 level for sample sizes under 50 across the number of independent variables, and regardless of sample size for 18 independent variables, in the case of OLS. In other words, the higher power rates for OLS and NB, particularly at the lower sample size values, come at the cost of inflated Type I error rates. With regard to the lasso, EN, and RR, for samples of 30 or more power rates exceeded 0.8 regardless of the number of independent variables, and was 0.85 or higher for RR when the sample size was 30 or more, regardless of the number of independent variables. Indeed, of these three regularization methods, RR displayed somewhat higher power rates across conditions simulated here. When the model included more independent variables, the gap between OLS and NB versus the regularization methods widened, particularly for smaller sample sizes. The exception to this pattern was for 18 independent variables and N=10, for which power rates were within 0.08 of one another. However, this gap between NB/OLS and the regularization estimators widened for *N*=20 and *N*=30, before narrowing for *N*=50 and *N*=100, for 18 independent variables. At the largest sample size condition, power for all of the methods were within 0.05 of one another, regardless of the number of independent variables.

Empirical Example

Coefficients for each of the DKEFS measures with respect to the verbal comprehension score appear in Table 3. The full set of results for all of the dependent variables are not presented here due to space limitations. However, they are available from the authors upon request. Given these results, it is clear that OLS had difficulty in obtaining coefficient estimates for the independent variables. Indeed, it only yielded estimates for the first 9 independent variables entered into the model, which was expected given that the total sample size was 10. The standardized regression values that were estimated appear to be quite unstable, given that their absolute values are typically 2 or more. At the other extreme in terms of estimation were the lasso and SS methods, each of which yielded coefficient values of 0 for each of the independent variables. In other words, both techniques shrunk the coefficient estimates to 0, suggesting that there were no relationships between the independent and dependent variables. Of the regularization approaches, RR produced the most non-0 coefficient estimates, and generally larger values when compared with EN or the lasso. Finally, when compared to OLS, NB yielded standardized coefficient values that appear to be more reasonable than those of OLS, though they differ from those of RR and EN in a number of respects. Of course, given that this is not simulated data, the true population values of these coefficients cannot be known. However, we can refer to the simulation results reported above in order to gain some insights into which technique's estimates are likely to be closest to the population values. Specifically, the simulation results for the 4 dependent variables 18 independent variables condition, which is most similar to the data structure in the empirical example, are useful in this regard. Based on the results in Table 1, we saw that though all methods yielded biased results in this combination of conditions, the least amount of such bias was present for the regularization methods, when compared to

Table 3. Standardized Coefficients for Empirical Data Analysis: Verbal Comprehension

Independent Variable	OLS	Lasso	EN	RR	NB	SS
Visual scanning	8.07	0	-0.001	-0.03	-0.001	0
Number sequencing	-7.11	0	0.01	-0.05	0.04	0
Letter sequencing	0.56	0	-0.001	0.03	-0.20	0
Numberletter sequencing	-3.33	0	0.15	0.16	0.06	0
Motor speed	-5.35	0	0.001	0.07	0.30	0
Letter fluency	9.22	0	0.01	0.11	0.08	0
Category fluency	1.87	0	0.02	0.12	-0.003	0
Category switching	2.10	0	0	-0.02	0.36	0
Category switching accuracy	-7.26	0	0	-0.01	-0.34	0
Filled dots	NA	0	0.003	0.11	-0.21	0
Empty dots	NA	0	0.14	0.30	-0.10	0
Dots switching	NA	0	0.10	0.11	1.16	0
Color naming	NA	0	0.02	0.05	0.35	0
Word reading	NA	0	0	-0.05	-0.20	0
Inhibition	NA	0	0	0.08	0.02	0
Inhibition/switching	NA	0	-0.03	-0.05	-0.03	0

NB. Thus, although we do not know the true values of the population parameters for the empirical data, given the findings from the simulation study, it seems likely that the RR estimates may be closer to the true value than are those for NB. Finally, then, it would seem that there are positive relationships between scores on the verbal comprehension scale and those on number-letter sequencing, letter fluency, category fluency, filled dots, empty dots, and dots switching.

Discussion

Researchers are frequently faced with the problem of high dimensional data, in which the sample size is relatively small, and there exist a relatively large number of variables in the statistical model of interest. In such cases, standard methods may not provide accurate parameter estimates or hypothesis test results (Bühlmann & van de Geer, 2011). In the context of univariate regression, regularization methods have been shown to provide more efficient estimates than standard OLS regression, with relatively little bias. The current study extends work in this area to the multivariate case, where high dimensionality might prove to be even more problematic than for univariate data (Bühlmann & van de Geer). In such situations, data analysts need access to statistical tools that can accommodate such potentially problematic data structure.

The simulation study reported above yielded promising findings for researchers faced with high dimensional multivariate data. In particular, the regularization methods, especially RR, yielded estimates that were somewhat more biased than OLS and NB for very small sample sizes, but which had lower standard errors and Type I error rates that were largely in control. The bias exhibited by the regularization methods and SS is to be expected, given that each of these approaches is expressly designed to suppress the parameter estimates (Tibshirani, 1996). And indeed, results similar to those reported here in the multivariate case, have also been shown with univariate regression (Zou & Hastie, 2005). The current results add to the literature by showing that this effect is somewhat magnified when more than 1 dependent variable is included in the analysis. However, it is also important to note that the when a model involves a large number of independent variables (e.g., 18), and a small sample size (e.g., 20 or fewer), the bias in OLS and NB estimates exceeds that for regularization methods and SS. Therefore, in such situations, these regularization approaches may be preferable.

Based on prior work with regularization estimators (Hoerl, 1962; Tibshirani, 1996; Zou & Hastie, 2005), it was expected that standard errors for the lasso, EN, and RR would be lower than for OLS or NB. And in fact, for all but the simplest models (i.e., those with 3 independent variables) the standard errors of lasso, EN, and RR estimates were consistently lower than those of OLS and NB, as was expected. Perhaps of more direct interest to applied researchers and data analysts, the Type I error rates for both

OLS and NB were well above the nominal 0.05 level whenever the sample size was 30 or fewer, and when it was 50 or fewer when there were 18 independent variables. This is of particular relevance because the Type I error rates will have a direct impact on research findings. Very practically, researchers using either OLS or NB with high dimensional multivariate data are more likely to conclude that one or more statistically significant relationships exist, when in fact they do not. Based on the findings presented above, when the data analyst is using multivariate regression, she must be concerned about this issue for as few as 3 independent variables, when the sample size includes 30 or fewer individuals, even when only 2 dependent variables are present in the model. It is also important to note that in these same conditions, the regularization approaches that control the Type I error rate also yield lower power than do either OLS or NB. The difference in power is most exaggerated for smaller sample sizes and more independent variables. In practice, then, data analysts will be faced with the choice of which potential error is more serious, identifying relationships that do not actually exist, or missing relationships that do. Obviously, the more serious error will depend upon the research scenario. Finally, if the only goal is control of the Type I error rate under all conditions, the SS estimator is the optimal method to use. It is certainly conceivable that a researcher would like to avoid making such an error above and beyond all else, if the consequence of such is relatively dire. In that instance, the SS estimator is clearly the preferred option, based on the simulation results presented above.

Limitations and Directions for Future Research

The goal of the current study was to extend the literature regarding regularization methods to include the multivariate case. It is hoped that this work has indeed done so. At the same time, it is important to acknowledge that there are limitations in the current study, and that further work in this area is still needed. For example, the number of independent variables included in the model was 3, 6, or 18. It would certainly be of interest to examine the performance of these methods for intervening values such as 12 or 15. In addition, the models used to generate the data were always linear, with no interactions or higher order terms included. Future work should include such terms, however. With respect to the Bayesian models, a wider array of prior distributions should be included in future research, in order to ascertain whether, and to what extent, such settings impact the performance of the methods, particularly with respect to parameter estimation accuracy. This issue is particularly trenchant for the small sample size cases, where the prior has a relatively large impact on the performance of the estimator (Kaplan, 2014). Finally, future work should examine the performance of these methods when the dependent variables do not come from a multivariate normal distribution. It would be of special interest to understand their performance when the outcome variables are skewed and/or kurtotic.

Conclusions

It is hoped that the results of this study will have direct application for researchers in practice. Based on the findings outlined above, we believe that there are some guidelines that may be helpful in this regard.

1. For high dimensional data, where the number of independent variables is close to, or exceeds the sample size, any of the three regularization methods (lasso, EN, and RR) are preferable to either OLS or NB.
2. Among the regularization methods, RR appears to yield the best tradeoff between Type I error control and power, making it a strong candidate for researchers to use in practice.
3. If primary interest is in the accuracy of parameter estimation and not control of the Type I error rate (i.e., inference is not as important as estimation accuracy) OLS or NB are preferable to the other methods studied here, except for cases in which the sample size is 20 or fewer and there are a large number (e.g., 18) of independent variables, or there are 6 dependent variables coupled with the small sample size. In such cases, the lasso, EN, and RR are all likely to yield estimates with much lower bias than OLS or NB.
4. If the only (or greatly overriding) concern is control of the Type I error rate, then the SS estimator is to be preferred.

References

Bradley, J. V. (1978). Robustness? *The British Journal of Mathematical & Statistical Psychology, 31*, 144-152.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Berlin: Springer-Verlag.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan executive function system*. San Antonio: Pearson.

Finch, W. H., Hernandez Finch, M. E., & Moss, L. (2014). Dimension reduction regression techniques for high dimensional data. *General Linear Model Journal, 40(2),* 1-15.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer-Verlag.

Hoerl, A.E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress, 58*, 54-59.

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, *33(2)*, 730-773.

Kaplan, D. (2014). *Bayesian statistics for the social sciences.* New York: The Guilford Press.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* Amsterdam: Elsevier.

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression standard errors and Bayesian lassos. *Bayesian Analysis, 5(2),* 369-411.

Lu, Z. H., Chow, S. M., & Loken, E. (2016). Bayesian factor analysis as a variable-selection problem: Alternative priors and consequences. *Multivariate Behavioral Research, 51(4),* 519-539.

McLean, R. L., Johnson, A., Zimak, E., Joseph, R. M., & Morrow, E. M. (2014). Executive function in probands with autism with average IQ and their unaffected first-degree relatives. *Journal of the American Academy of Child and Adolescent Psychiatry, 53(9),* 1001-1009.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling, 23(5),* 750-773.

R Core Development Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rockova, V., & George, E. I. (2016). The Spike-and-Slab LASSO. *Journal of the American Statistical Association* (accepted).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B., 58*, 267-288.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition*. San Antonio: Pearson.

Xu, X., & Shosh, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis, 10(4),* 909-936.

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

Send correspondence to: W Holmes Finch
Ball State University
Email: whfinch@bsu.edu