Biostatistics
Sample Work

Techniques and applications for regularization methods in clinical biostatistics

Abstract

In the data sciences, a variety of regularization algorithms have been proposed to overcome overfitting, leverage sparsity, or enhance prediction. We discuss a variety of techniques within this framework, including penalization, early halting, ensembling, and model averaging, using a wide definition of regularization, which involves regulating model complexity by adding information in order to solve ill-posed problems or prevent overfitting. Aspects of their actual implementation are explored, as well as accessible R-packages and examples. We surveyed three general medical publications to determine the extent to which these techniques are employed in medicine. With the exception of random effects models, it demonstrated that regularization procedures are rarely used in real clinical applications. As a result, we propose that regularization procedures be used more frequently in medical research. The sole disadvantage of regularization procedures in instances when other approaches work well is increased complexity in the conduct of the biostatistics studies, which might provide obstacles in terms of computer resources and skill on the part of the data analyst. Both can and should, in our opinion, be addressed by investing in proper computing infrastructure and instructional resources.

Keywords:

Penalization, Bayesian inference, ensembling, model averaging, early stopping, evidence synthesis

INTRODUCTION

Regularization's overarching goal is to control model complexity by providing information, allowing us to solve ill-posed problems and avoid overfitting. Regularization strategies include penalization,1,2 early stopping,3,4 ensembling5,6 and model averaging within this wide definition.7 These statistical tools have long been used in medical research. Penalization, for example, is used in variable or model selection using ridge regression1 or the least absolute shrinkage and selection operator (LASSO).2 These methods can also be used in the context of missing data8, causal analyses9, and so on. Bayesian hierarchical models are also utilized for evidence synthesis.10 Whereas standard meta-analysis focuses on the cumulative impact across a number of included studies, the same hierarchical models may also be used for dynamic borrowing, which is the estimation of an effect in one research using information borrowed from previous studies via shrinkage estimation.11 Regularization has clinical uses ranging from pharmacovigilance12 to non-small-cell lung cancer13 to Alzheimer's disease.14

Although there is a growing literature on regularization and a multitude of strategies available to solve the challenges listed above, it is currently unknown to what degree these methods are employed in clinical care and what kind of problems they address. To help answer these issues, we conducted a systematic evaluation of recent volumes of three general medical journals: the Journal of the American Medical Association (JAMA), the New England Journal of Medical (NEJM), and the British Medical Journal (BMJ).

The rest of this paper is structured as follows. Section 2 provides an overview of regularization procedures, beginning with a brief history of regularization and focusing on topics such as penalization, early halting, ensembling, and model averaging. Section 3 reports on a survey of papers from medical journals that outline the present state of regularization applications in clinical medicine. Section 4 has some instances before Section 5 concludes with some final remarks.

2.REGULARIZATION APPROACHES

In this part, we will discuss several regularization methods. We will develop precise goals as well as appropriate statistical models and strategies to achieve them. Penalization and incorporating external and historical data (Section 2.1), early halting (Section 2.2), ensembling (Section 2.3), and other concepts such as introducing noise (Section 2.4) are examples of regularization procedures. Table 1 outlines all of these regularization kinds, their aims, and the statistical approaches used to achieve them. This part finishes with some practical observations on regularization (part 2.6) and related software (Section 2.7).

Type	Description	Common statistical approaches
Penalization (Section 2.1)	Add penalty term(s) to fitting criterion	– Ridge regression, LASSO, elastic net – Bayesian regularization priors – Constraints for parameters – Random effects – Semiparametric regression
Early stopping (Section 2.2)	Early stopping of an iterative fitting procedure	– Coefficient paths in penalization approaches – Boosting – Pruning of trees – Learning rate in deep neural networks
Ensembling (Section 2.3)	Combine multiple base-procedures to an ensemble	– Bagging – Random forests – (Bayesian) model averaging – Boosting
Other approaches (Section 2.4)	–	– Injecting noise – Random probing in model selection – Out-of-sample evaluation

PENALIZATION

By combining (a) a (lack of) fit criteria expressing a model's capacity to match the provided data with (b) a penalty that quantifies model complexity, penalization techniques make the trade-off between model fit and model complexity clear. This approach will be introduced in greater depth in the following for parametric models with parameter vectors, although the principles are instantly generalizable to semi- and nonparametric models. The letter y will indicate the observed data, and we will demonstrate penalization along regression-type models, where y represents a vector of recorded response values and contains the regression coefficients.

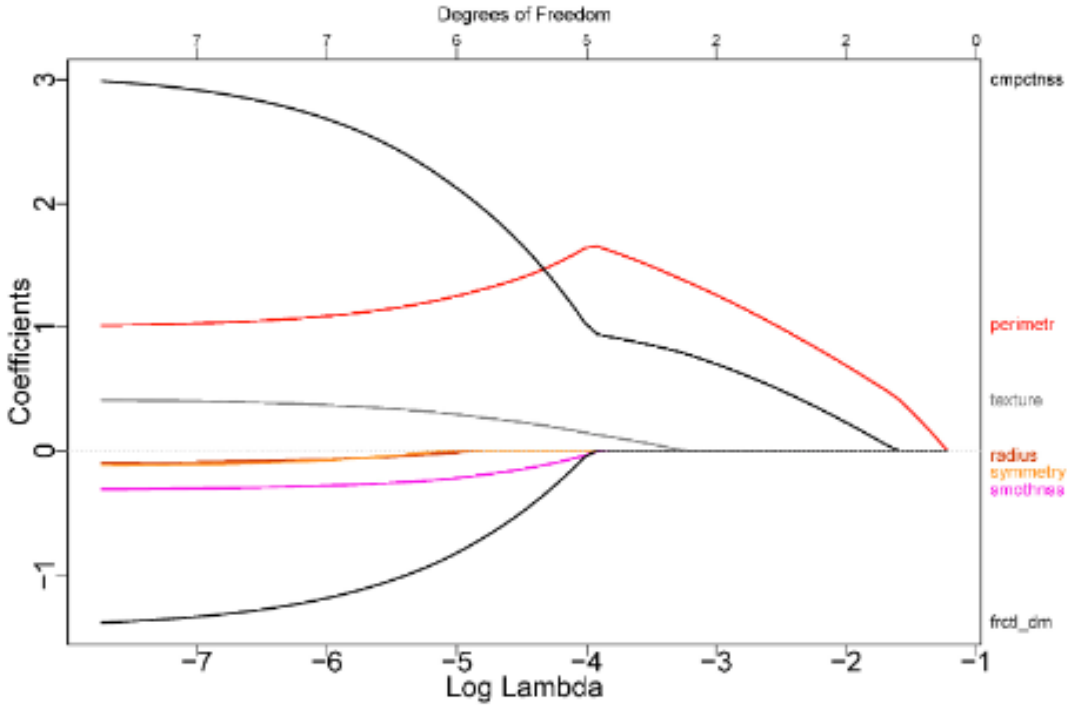

Figure 1. Coefficient paths obtained by applying the least absolute shrinkage and selection operator (LASSO)

Finally, applying the penalty to functions rather than the parameter vector allows for the enforcement of various sorts of regularization behaviour. Furthermore, including the penalty on transformations or basis function expansions of the original covariates adds flexibility. Some areas of significant focus in the recent decade have been:

•*Fusion penalties:* When evaluating the impacts of characteristics that may be arranged in any meaningful way, the objective is to fuse particular effects. Ordinal categorical covariate effects are just one example of this. The 'fused LASSO'33, which penalizes the L1 norm of both coefficients and their subsequent differences, was one of the early recommendations, although numerous extensions have been proposed in the literature since then.34-39 Tutz and Gertheiss provide a comprehensive analysis of the topic 'Regularized regression for categorical data' for both categorical predictors and responses.40 Several scholars have examined effect fusion within the Bayesian framework, including Pauger et al.41 and Malsiner-Walli et al.42.

•*Semiparametric function estimation* Using smoothness priors where a flexible influence of a covariate of interest must be evaluated. Working with function spaces and associated norms, such as the functional L2 loss pen, which is the integrated squared second derivative that penalizes function curvature, is one way. This is the foundation for the well-known particular case of smoothing splines. When estimating the impact of interest in terms of a basis expansion, penalties can be created for the basis coefficients using penalized splines44,45, being one of the most popular examples. Penalties may, therefore, be designed to enforce not just smoothness but also additional features such as monotonicity, convexity/concavity, and continuous limiting behaviour.46,47

•*Structured additive regression models* that evaluate regression predictors that are an additive mixture of several forms of effects based on covariate vectors of diverse types and connected with quadratic penalties to enforce desired qualities of the individual effects. Structured additive regression, for example, includes nonlinear effects of continuous variables, changing coefficient terms, interaction surfaces, random effects, and spatial effects as special examples; for further information, see Fahrmeir et al.48 and Fahrmeir and Kneib49.

•*Single index models* that, in a data-driven manner, expand generalized linear and additive models by estimating the link function that translates the regression predictor to the conditional expectation of the response variable. Regularization is necessary for this section of the model when a flexible, nonparametric method is used for the link function.50 Single index models with a linear predictor combine nonlinear and linear modelling approaches.

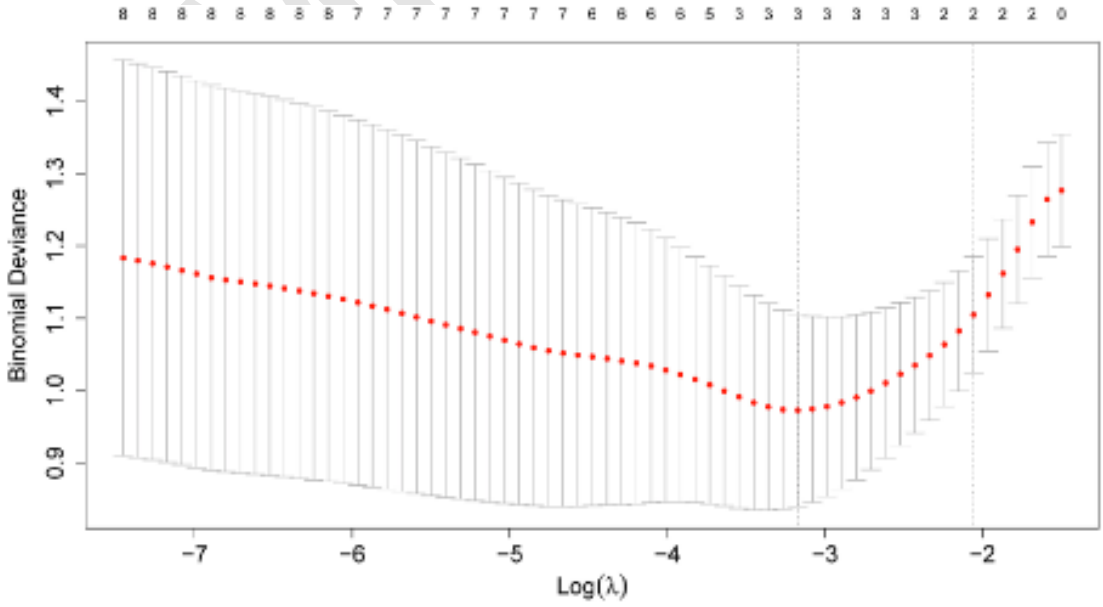

Figure 2. Cross-validation error curve for the LASSO applied to the prostate cancer data from Section 2.7.

Early stopping

Many statistical and machine learning algorithms construct a (possibly complicated) model by iteratively improving a basic model toward the maximum complex scenario permitted by the model specification. In such cases, one method of inducing regularization is to stop the fitting process before the most complex model is achieved, i.e. to identify the best trade-off between model simplicity (models close to the initial model) and fit the data (models close to the final, most complex model) by stopping early. Indeed, when considering the complete path of coefficients produced by varying the penalty parameter from infinity (simplest model determined by minimizing the penalty) to zero (complex model fit without penalization), the penalization approaches discussed in the previous section can also be cast into this framework.

Ensembling and model averaging

While the previous two techniques explicitly included regularization into a single model, we now shift to regularization by mixing a number of models with the goal of improving model performance. Consider a model that has a high capacity to match the provided data but a high variability, such that the model does not generalize well to new data. If numerous variations of such a model are available, the variability can be decreased by assembling the models into an ensemble or averaging across predictions or other values produced from the models.

Other regularization approaches

Regularization approaches are various methods used to improve model-fitting processes. Some examples include injecting noise, random forests, random probing, out-of-sample evaluation strategies, and drop-out in neural networks. Injecting noise introduces distortion in the model-fitting process, while random forests have two steps of randomization. Random probing introduces simulated additional covariates independent of the response of interest, helping distinguish informative and non-informative covariates in model selection procedures. Out-of-sample evaluation strategies determine the model's ability to generalize beyond observed data based on hold-out datasets. Drop-out in neural networks randomly shuts down part of neurons in one layer to avoid overfitting due to co-adaption.

R implementation of different regularization approaches

The extra material of this work contains an example implementation of several regularization algorithms. For example purposes, we utilized a Kaggle data set on prostate cancer.79 The data includes information on 100 patients' tumours (radius, texture, perimeter, and so on) as well as their diagnosis (binary result). We compare the area under the curve (AUC) and mean classification error (MCE) of six different regularization approaches, namely a classification tree (CART), a random forest, subset selection, rigde regression, LASSO, and elastic net, to standard logistic regression. The hyperparameters are selected using a 10-fold CV, and the results are averaged across ten repeats. The regularization approaches surpass ordinary logistic regression in this scenario.

The state of regularization applications in medicine

We surveyed the literature in three prominent medical journals to determine how much regularization is utilized in published medical research. To that end, we examined all issues of the Journal of the American Medical Association (JAMA), the New England Journal of Medicine (NEJM), and the British Medical Journal (BMJ) published between January and September 2020. These publications were chosen because they are among the top general medical journals in the world in terms of impact factors. We identified and examined all original research publications, yielding 383 articles; see the PRISMA flow chart in the appendix for more information.

Overview of used regularization methods

The study screened 380 articles for regularization applications using statistical methods, using the definition and examples provided in Section 2. The results were collected in an Excel spreadsheet, which includes the digital object identifier, journal, first author's name, title, sample sizes, and the type of software used for analyses. The study also extracted the type of software used for the analyses. The main findings regarding the use of regularization are summarized in Table 2, with an additional table summarizing study characteristics. The results were collected in an Excel spreadsheet, which is provided as a supplement.

	No regularization	Random effects	Bayes	Penalization	A priori	CV	Smoothing	Boosting	Random forest	Subset selection
JAMA	62 (61%)	35 (35%)	1 (1%)	2 (2%)	0 (0%)	1 (1%)	0 (0%)	1 (1%)	1 (1%)	0 (0%)
NEJM	121 (74%)	31 (19%)	8 (5%)	1 (1%)	2 (1%)	1 (1%)	2 (1%)	0 (0%)	0 (0%)	0 (0%)
BMJ	70 (60%)	38 (33%)	7 (6%)	3 (3%)	1 (1%)	0 (0%)	1 (1%)	1(1%)	0 (0%)	1 (1%)
Total	253 (67%)	104 (27%)	16 (4%)	6 (2%)	3 (1%)	2 (1%)	3 (1%)	2 (1%)	1 (0%)	1 (0%)

CV: cross-validation; JAMA: Journal of the American Medical Association; NEJM: New England Journal of Medicine; BMJ: British Medical Journal.

Discussion of specific examples

Other than random effects modelling, each regularization approach described in Table 2 is briefly discussed in the examined works. We do not include these works in the references in the rest of this section since they act as examples rather than literature citations. The online supplement contains all of the necessary information.

Examples

We review chosen biostatistical examples from the literature to highlight the adaptability of applying regularization approaches and their possible positive impacts.

Variable selection and shrinkage methods for linear regression

The prostate cancer data set examined in Chapter 3.4 of Hastie et al.87 with shrinkage approaches for linear regression is a well-known example in the statistical learning field. The information comes from research conducted by Stamey et al.88, who evaluated the level of prostate-specific antigen (PSA) in 97 prostate cancer patients prior to radical prostatectomy. The correlation of log PSA (lpsa) to eight clinical variables was investigated, including log cancer volume (lca), log prostate weight (weight), age, log of benign prostatic hyperplasia amount (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).

Using additional a priori information for evidence synthesis

Borrowing information, such as from observational research, to support a small-scale randomized trial can be accomplished through the use of a shrinkage estimate within a Bayesian random effects meta-analysis.11 The method first analyzes the observational data in a hierarchical model with the shrinkage estimator and then uses the resultant posterior distribution to inform the RCT analysis. Röver and Friede11 demonstrated the efficiency benefit of this strategy in the setting of Creutzfeldt-Jakob illness. This is an uncommon illness with an incidence of 1 in 1,000,000 people. An RCT on doxycycline89 was halted early, with just 12 patients participating. An observational research, however, provided further data on 88 individuals. Cox proportional hazards regression was used to examine the primary endpoint of all-cause death.

Boosting capture-recapture methods

Systematic reviews of clinical trials should include all relevant studies on the subject. Capture-recapture analyses have been developed to assess the comprehensiveness of systematic literature reviews. These need the selection of a suitable model. Rücker et al.92 proposed combining capture-recapture analysis with componentwise boosting to achieve this goal. The boosting technique allows you to define both necessary and optional variables that are always included in the model. The latter are only included if they are relevant. This method proved to be resistant to overfitting, and an effective model for statistical inference was generated automatically. Rücker et al.92, in particular, compared componentwise boosting to a manually chosen Poisson model to predict the number of missing references for two systematic reviews. The manually chosen model predicted 82 missing articles (95% CI: 52-128) in the first analysis, whereas the boosting approach discovered 127 (95% CI: 86-186) missing articles, and in the second case, boosting produced a more efficient estimate of 188 (95% CI: 159-223) than the best manually selected model (140 missing articles with a 95% CI: 116-168).

Conclusion

A variety of regularization algorithms have been proposed to address issues such as overfitting, data sparsity, and improving prediction and generalizability of outcomes. We explored a variety of techniques within this framework, including penalization, early halting, ensembling, and model averaging, using a wide definition of regularization, namely the act of adding information to regulate model complexity. We talked about the practical issues of their implementation, such as the various R-packages. In this paper, we focused on R as a programming language and showed how to employ regularization methods in an R implementation. Regularization procedures, on the other hand, are also incorporated in other statistical tools. Penalization methods such as LASSO and Ridge regression, for example, are implemented in SAS's GLMSELECT and REG procedures, whereas more complicated penalization methods may be found in PROC TPSPLINE. PROC HPFOREST, for example, provides a random forest implementation. The Bayesian technique may be used in a variety of ways, including PROC FMM, PROC GENMOD, PROC LIFEREG, and PROC PHREG support Bayesian analysis via the BAYES statement, whereas PROC BGLIMM and PROC MCMC are specially designed for Bayesian estimation.93 Similarly, xtreg, lasso, and boost in Stata implement random effects models, LASSO penalization, and boosting, respectively. Examples were supplied to demonstrate the actual application of regularization in order to encourage more widespread adoption of these techniques in medicine. This is on the background of our review of recent issues of three general medical journals, which revealed that regularization approaches could be used more. The main exception is random effects models, which appear very often. Other regularization methods were rarely used. We believe that there is room for improvement in the application of regularization approaches in clinical medicine. They may be used on a frequent basis because they only improve analysis and interpretation. The sole disadvantage of regularization procedures in instances when other approaches work well is increased complexity in the conduct of the studies, which might provide obstacles in terms of computer resources and skill on the part of the data analyst. Both can and should, in our opinion, be addressed by investing in proper computing infrastructure and instructional resources.

REFERENCES

1.Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970; 12: 55–67. Crossref. ISI.

2.Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996; 58: 267–288. Crossref.

3.Mayr A, Binder H, Gefeller O, et al. The evolution of boosting algorithms - from machine learning to statistical modelling. Methods Inf Med 2014; 53: 419–427. Crossref. PubMed. ISI.

4.Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting (with discussion). Stat Sci 2007; 22: 477–505. Crossref. ISI.

5.Breiman L. Bagging predictors. Mach Learn 1996; 24: 123–140. Crossref. ISI.

6.Breiman L. Random forests. Mach Learn 2001; 45: 5–32. Crossref. ISI.

7.Claeskens G, Hjort NL. Model selection and model averaging. Cambridge, UK: Cambridge University Press, 2008. Crossref.

8.Ch Tseng, H Chen Y. Regularized approach for data missing not at random. Stat Methods Med Res 2019; 28: 134–150. Crossref. PubMed.

9.Ye Z, Zhu Y, Coffman DL. Variable selection for causal mediation analysis using LASSO-based methods. Stat Methods Med Res 2021; 30: 1413–1427. Crossref. PubMed. ISI.

10.Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and healthcare evaluation. vol. 13. Chichester: John Wiley & Sons, 2004.

11.Röver C, Friede T. Dynamically borrowing strength from another study through shrinkage estimation. Stat Methods Med Res 2020; 29: 293–308. Crossref. PubMed. ISI.

12.Ahmed I, Pariente A, Tubert-Bitter P. Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. Stat Methods Med Res 2018; 27: 785–797. Crossref. PubMed. ISI.

13.Gross O, Tönshoff B, Weber LT, et al. A multicenter, randomized, placebo-controlled, double-blind phase 3 trial with open-arm comparison indicates safety and efficacy of nephroprotective therapy with ramipril in children with Alport's syndrome. Kidney Int 2020; 97: 1275–1286. Crossref. PubMed.

14.Kruegel J, Rubel D, Gross O. Alport syndrome–insights from basic and clinical research. Nat Rev Nephrol 2013; 9: 170. Crossref. PubMed.

15.Rücker G, Reiser V, Motschall E, et al. Boosting qualifies capture–recapture methods for estimating the comprehensiveness of literature searches for systematic reviews. J Clin Epidemiol 2011; 64: 1364–1372. Crossref. PubMed.

16.Website of SAS Institute. https://sas.com. Accessed on June 29th, 2022.

17.Riley RD, Snell KI, Martin GP, et al. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. J Clin Epidemiol 2021; 132: 88–96. Crossref. PubMed.

18.Sauerbrei W, Abrahamowicz M, Altman DG, et al. Strengthening analytical thinking for observational studies: the stratos initiative. Stat Med 2014; 33: 5413–5432. Crossref. PubMed.

19.Sauerbrei W, Perperoglou A, Schmid M, et al. State of the art in selection of variables and functional forms in multivariable analysis-outstanding issues. Diagn Progn Res 2020; 4: 3. Crossref. PubMed.