Journal of Data Science 19 (1), 1–14
January 2021
DOI: 10.6339/21-JDS995
Computing in Data Science

A Simple Aggregation Rule for Penalized Regression Coefficients after Multiple Imputation

Ryan A. Peterson[1,*]

[1]*Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado-Denver Anschutz Medical Campus, Aurora, Colorado, USA*

Abstract

Early in the course of the pandemic in Colorado, researchers wished to fit a sparse predictive model to intubation status for newly admitted patients. Unfortunately, the training data had considerable missingness which complicated the modeling process. I developed a quick solution to this problem: Median Aggregation of penaLized Coefficients after Multiple imputation (MALCoM). This fast, simple solution proved successful on a prospective validation set. In this manuscript, I show how MALCoM performs comparably to a popular alternative (MI-lasso), and can be implemented in more general penalized regression settings. A simulation study and application to local COVID-19 data is included.

Keywords *elastic net; LASSO; minimax concave penalty; missing data; regularization*

1 Introduction

Aggregating results after multiple imputation is not always straightforward, especially in circumstances where more complex analyses are necessary on each imputed data set. Such is the case for the use of penalized regression to simultaneously estimate and select from a set of candidate features. While some solutions have been proposed for this question, existing methods do not necessarily have robust software for quick and efficient implementation. Amid the coronavirus disease 2019 (COVID-19) pandemic, speed of analysis has been relatively important; statisticians are under pressure to get useful information into the hands of clinicians, health care workers, and policy makers as soon as possible in order to prevent future cases and optimally utilize the knowledge and data we have available. However, balancing this need for speed while retaining analysis quality can be challenging. A recent systematic review has shown that while numerous predictive models for COVID-19 have been developed, virtually all of them are subject to various flaws and biases (Wynants et al., 2020), in many cases stemming from improper treatment of missing data.

Soon after University of Colorado Hospital (UCH) started seeing severe COVID-19 patients requiring intubation, some investigators wished to perform a data analysis on these first patients to see if patient data upon admission could predict whether the patient would require intubation or not. Their goal was to have an understandable, interpretable formula for the risk of intubation for any particular patient entering the hospital with COVID-19. Unfortunately, the data exhibited a fair amount of missingness which was clearly not completely random, so the primary statistical task became to build a sparse predictive model for the risk of intubation in the presence of missing data. I was not the only one to be confronted with this problem; Gong et al. (2020) developed a similar sparse model to predict severe COVID-19 in the presence of missing data, however they utilized only a single imputation step as opposed to multiple imputation.

* Email: ryan.a.peterson@cuanschutz.edu.

Received July, 2020; Accepted October, 2020

With speed of analysis being paramount, I developed and implemented an ad hoc method for aggregating penalized regression coefficients that I call *Median Aggregated penaLized Coefficient after Multiple imputation* (MALCoM). Instead of aggregating models across multiple imputations using Rubin's Rules (Rubin, 2004) for pooling means and standard errors, MALCoM simply fits a penalized regression model on each imputed data set separately and takes the median of each coefficient across imputations. Taking the median instead of the mean ensures the final model is sparse should variable selections be inconsistent across imputations.

In this work, I go over several similar aggregation methods, and compare the predictive and selective performance of MALCoM relative to a major competing framework in a series of simulations. I then use both methods on local COVID-19 data from the UCH to build prediction models for intubation and compare their performance on a prospective validation cohort. I conclude with a discussion of strengths and weaknesses of MALCoM relative to other methods.

2 Methodology

The primary outcome $\boldsymbol{y}$ is defined as of a vector of n Bernoulli random variables, where $y_i = 1$ if patient i was intubated, and $y_i = 0$ if they were not. Also, X denotes a $n \times p$ covariate matrix where we suspect some of the p columns relate to the probability of intubation, and others do not. The ultimate goal is to build a prediction equation to easily calculate intubation risk given a (hopefully sparse) set of columns of X.

Building such a risk equation via penalized regression when either the $\boldsymbol{y}$ vector or the X matrix is incomplete/missing is not trivial. When the data are missing completely at random (MCAR), list-wise deletion (sometimes called complete-case analysis) is not a biased approach to analysis. However, many observations with informative power about the patterns present in the data get thrown out entirely, which can decrease precision considerably (and unnecessarily). Further, the MCAR assumption is unrealistic in many if not most circumstances. Multiple imputation, on the other hand, is not only valid when data are missing at random (MAR), but the procedure can also benefit from increased precision resulting from the use of all of the observed data. Further, multiple imputation is preferred to single imputation because it is generally better able to account for uncertainty resulting from the imputation process.

Therefore, this work focuses on how to use penalized regression methods such as the lasso (Tibshirani, 1996), the elastic net (Zou and Hastie, 2005), and the minimax concave penalty (MCP) (Zhang, 2010) in combination with settings where multiple imputation is warranted. In this section, I will provide a brief overview of the types of penalized regression, introduce the MALCoM method as a simple, effective and versatile approach, and describe other existing methods for aggregating penalized regression models after multiple imputation.

2.1 Penalized Logistic Regression

I will henceforth treat the question at hand using a linear model, where $\boldsymbol{\beta}$ refers to the regression coefficients, and if $\beta_j = 0$, this indicates covariate j (the j^{th} column of X) is not related to the probability of success ($\boldsymbol{\pi}$) for $\boldsymbol{y}$.

The objective function for penalized logistic regression is:

$$Q(\boldsymbol{\beta}|X, \boldsymbol{y}) = -\frac{1}{n}\sum\{y_i \log \pi_i + (1 - y_i)\log(1 - \pi_i)\} + P(\boldsymbol{\beta}|\boldsymbol{\theta}),$$

where $\boldsymbol{\pi} = \exp(X\boldsymbol{\beta})/(1 + \exp(X\boldsymbol{\beta}))$ and $P(\boldsymbol{\beta}|\boldsymbol{\theta})$ refers to a penalty on the magnitude of the

coefficients in $\boldsymbol{\beta}$ (the hyperparameter(s) denoted here by $\boldsymbol{\theta}$ determine the extent to which coefficients are shrunk/penalized). For the lasso, $P(\boldsymbol{\beta}|\lambda) = \lambda \sum_j |\beta_j|$. For MCP, an additional tuning parameter controls the concavity of the penalty, and the penalty function is given by

$$P(\beta_j|\lambda, \gamma) = \begin{cases} \lambda|\beta_j| - \beta_j^2/2\gamma, & \text{if } \beta_j \leqslant \gamma\lambda \\ \gamma\lambda^2/2, & \text{otherwise} \end{cases}$$

and $P(\boldsymbol{\beta}|\lambda, \gamma) = \sum_j P(\beta_j|\lambda, \gamma)$. For elastic net, another tuning parameter α controls the balance of penalty attributed to the lasso penalty and the ridge penalty: $P(\boldsymbol{\beta}|\lambda, \alpha) = \alpha\lambda \sum_j |\beta_j| + \frac{(1-\alpha)\lambda}{2} \sum_j \beta_j^2$.

Comprehensive software to optimize each of these objective functions is available in R (Friedman et al., 2010; Breheny and Huang, 2011), although this software assumes complete response and covariate data. Therefore, when any amount of missingness exists, observations must be deleted in a list-wise fashion or imputed. For reasons mentioned previously, multiple imputation procedures, such as those implemented via chained equations via the `mice` software package in R (Van Buuren and Groothuis-Oudshoorn, 2011), have become popular. In these procedures, the analyst obtains M "augmented" data sets wherein data that had been missing in $\boldsymbol{y}$ and X have been imputed based on a series of (possibly complex) models. The analysis then can be performed M times on each of the imputed data sets and pooled to produce a final model and accompanying predictions/selections.

I will refer to the imputed response and covariate data as $\boldsymbol{y}^{(m)}$ and $X^{(m)}$ respectively, where $m \in \{1, 2, ..., M\}$ refers to the imputation index for each augmented data set. In the ordinary regression context, analyses will yield a series of estimates for $\hat{\boldsymbol{\beta}}^{(m)}$ (and their associated standard errors) which can be pooled using Rubin's Rules. In the penalized regression setting however, it is expected that some of these estimated coefficients will be exactly zero, thereby deselecting their associated covariates in the model. Model pooling is therefore complicated by the non-continuous nature of these estimates, as well as the fact that their standard errors are difficult to estimate. So, the primary question can be described: In the penalized regression setting, what is the best way to aggregate $\hat{\boldsymbol{\beta}}^{(m)}$ across M imputed data sets to build a single (hopefully sparse) prediction equation?

2.2 Median Aggregation of Penalized Coefficients after Multiple Imputation

If the analyst were to aggregate coefficients using the mean across imputations (i.e. $\hat{\beta}_j = \frac{1}{M} \sum_m \hat{\beta}_j^{(m)} \ \forall \ j$) in an ordinary regression context, no issues arise. However, when certain coefficients can be set to exactly zero, selected (nonzero) coefficients may differ across imputed data sets (e.g. $\hat{\beta}_j^{(1)} \neq 0$ but $\hat{\beta}_j^{(m)} = 0 \ \forall \ m \neq 1$). This inconsistency of selection combined with a mean aggregation rule ultimately yields models that are unnecessarily saturated, since the mean coefficient across data sets will be nonzero if any single one is nonzero.

The goal of taking the mean coefficient across the M coefficient sets is to determine the central tendency of the coefficient, and another simple measure of central tendency is the median. Using the median penalized coefficient across imputed data sets ensures that at least a majority of the M analyses must select a nonzero coefficient (in the same direction) in order for that coefficient to be nonzero post-aggregation. This idea forms the basis for MALCoM. Since the concept and implementation is simple, it is easy to implement quickly within existing software for fitting lasso/MCP/elastic net models. I therefore chose to use MALCoM in the original data

analysis. However, other methods exist that can successfully perform penalized regression in the context of multiple imputation, and I will briefly describe these in the next section.

2.3 Alternative Approaches

A detailed overview of model selection in the presence of missing data is outside the scope of this work, but useful overviews are available (Van Buuren, 2018; Zhao and Long, 2017). The Multiple Imputation lasso (MI-lasso) is a close and popular competitor to MALCoM whose main idea is to ensure that the $\boldsymbol{\beta}^{(m)}$ selections are consistent (i.e. the same set of covariates are selected) across all M data sets, so that a simple average can be taken across imputations Chen and Wang (2013). This goal of having a common set of variables selected across imputations is referred to in Chen and Wang (2013) as "selection consistency", however this is not to be conflated with the notion of asymptotic selection consistency (e.g., Sirimongkolkasem and Drikvandi, 2019; Yang and Yang, 2018). MI-lasso accomplishes its consistency by treating the modeling process for all augmented data sets as a group lasso problem (Meier et al., 2008), where each covariate j across the M data sets is treated as a single group, and a full model is fit using all data sets simultaneously. This grouping scheme ensures that if a coefficient j is selected to be nonzero for one augmented data set, all other coefficients for the other data sets will also be nonzero. Although unlikely, it is possible that the sign of the coefficient could differ across imputed data sets. The MI-lasso procedure can be implemented via a software program in R (available at http://www.columbia.edu/~qc2138/Downloads/software/MI-lasso.R) which will select an optimal λ via the Bayesian Information Criterion (BIC). Only the lasso penalty for Gaussian responses can be implemented with currently available software.

The Multiple Imputation Random Lasso (MIRL) is a more involved procedure that uses the random lasso and stability selection to produce more accurate predictions and variable selections across imputations (Liu et al., 2016). The MIRL leverages the power of bootstrapping and importance rankings and can be effective even in scenarios with a high proportion missing and where $p > n$. Unfortunately, at the time of writing, the primary software package that executes the MIRL was unavailable on the Comprehensive R Archive Network, and similar to the MI-lasso, its now archived implementation is only available for the lasso penalty and Gaussian responses. Another very similar method which uses bootstrapping in conjunction with stability selection is called the BI-SS (Long and Johnson, 2015), though it is still unknown how the MIRL and the BI-SS compare to one another (Van Buuren, 2018). Both of these methods, while optimal under certain situations, are computationally intensive.

Finally, an additional approach would be to produce outcome predictions for models produced on all of the multiply imputed data sets, and simply averaging these predictions to obtain a final prediction. In this ensemble approach, whether the variables selected across imputations are consistently the same is not of consequence. Indeed, when prediction accuracy is the only goal of the analysis, such an ensemble technique would likely work very well. However, when prediction is not the sole goal, and/or when those making predictions require a single sparse model, the parameters of the model themselves are quite important. For instance, in the application, the model was built for health care workers to triage patients based on intubation risk. The process of implementing 20 prediction equations each time a patient enters the hospital then averaging these results would be considerably more difficult than simply using the aggregated logit parameters to make one single prediction. Further, due to the instability of the model selection across multiply imputed data sets, these 20 prediction equations would likely require collecting additional covariates on these patients, which would lead to increased cost and time

spent unnecessarily. A single set of sparse coefficients, such as those produced by MALCoM, can often be more easily and frugally implemented.

3 Simulation

3.1 Simulation Design

This section presents a series of simulations showing how well MALCoM performs compared to the MI-lasso and the lasso applied to a full (nonmissing) data set. Simulations were grounded in what was observed in the UCH COVID-19 data. In particular, I used the following parameters:

- $n = 158$
- $p \in \{10, 40\}$
- Missingness mechanism: MAR
- X is multivariate normal with exchangeable correlation of $\rho = .45$ (setting 1), or multivariate normal with AR(1) correlation of $\rho = .1$ (setting 2)
- The proportion missing $p_{miss} \in \{0.3, 0.7\}$
- The signal-to-noise ratio $SNR \in \{0.3, 1.0, 3.0\}$
- The number of true nonzero effects $s = 3$, alternating positive and negative of equal magnitude
- The `mice` package is used to perform multiple imputation with ($M = 20$), and otherwise default settings.

Altogether this represents 24 unique parameter combinations. Admittedly there are many other ways of varying these parameters, and only a subset of them are able to be presented in this work. To mitigate this, I have included the simulation code as a supplemental script that allows custom specification of each of these parameters. An alternative simulation study that treats the response $\boldsymbol{y}$ as Gaussian is included in the supplemental as well.

MALCoM models were fit using several candidate approaches; I used lasso, MCP, and elastic net penalties (which will henceforth be abbreviated as MCL, MCM, and MCE respectively). In the original analysis, I used MALCoM combined with MCP because MCP tends to select a sparser model which is easier to interpret and implement in practice in a risk score, especially when some covariates have a tendency to be missing. The `ncvreg` R package was used to fit these models (Breheny and Huang, 2011). For MCP models, the tuning parameter γ was set to 3 (the default in `ncvreg`), while for elastic net models, the proportion of the overall penalty attributed to the L1 penalty (α) was set to 0.25 in order to promote a divergence in results between the elastic net and the lasso. In all models, λ was selected via 10-fold cross-validation (CV). Since MI-lasso has available software for only Gaussian responses, I used it to build a linear probability model in order to compare it with MALCoM. The linear probability model works directly to predict the outcome $\boldsymbol{y}$ and uses (penalized) least-squares to estimate model coefficients, which refer to changes in the probability of response directly rather than a change in the log odds. For the MI-lasso, λ was selected by BIC, since this is recommended and implemented in the program provided by the authors of this method.

For each combination of simulation parameters, 250 training data sets were simulated. For each of these, I fit all models described above and computed the area under the receiver operating characteristic (ROC) curve (AUC) on a newly generated data set of 1500 observation as a measure of predictive accuracy (Robin et al., 2011). I also calculated the false discovery rate (FDR) and the false negative rate (FNR) of the models' selections as an assessment of variable selection performance. To efficiently aggregate results across simulations, these performance metrics (AUC, FDR, and FNR) are then modeled using a linear mixed model with a

Table 1: Results from linear mixed model for AUC. Intercept refers to the expected AUC for a full-data lasso model tuned via 10-fold cross-validation. Beta coefficients represent the expected change in AUC (percentage points) if the corresponding method is used relative to the full-data lasso predictions (a coefficient of 0 indicates no loss of predictive performance when data are missing).

Missing Rate	p	SNR	Intercept	beta.MCL	beta.MIL
Low-missing, low-p					
0.3	10	0.3	59.19 (58.6, 59.7)	0.11 (−0.3, 0.5)	−1.93 (−2.3, −1.6)
0.3	10	1.0	70.93 (70.5, 71.3)	−0.19 (−0.4, 0)	−1.2 (−1.4, −1)
0.3	10	3.0	82.31 (82.1, 82.5)	−0.14 (−0.2, 0)	−0.22 (−0.3, −0.1)
Low-missing, high-p					
0.3	40	0.3	56.57 (56, 57.1)	0.24 (−0.1, 0.6)	0.23 (−0.1, 0.6)
0.3	40	1.0	68.65 (68.2, 69.1)	−0.06 (−0.3, 0.2)	−0.55 (−0.8, −0.3)
0.3	40	3.0	80.94 (80.7, 81.2)	−0.06 (−0.2, 0.1)	0.13 (0, 0.3)
High-missing, low-p					
0.7	10	0.3	58.8 (58.3, 59.3)	0.18 (−0.2, 0.6)	−1.64 (−2, −1.3)
0.7	10	1.0	71.06 (70.7, 71.4)	−0.38 (−0.6, −0.2)	−1.23 (−1.5, −1)
0.7	10	3.0	82.24 (82, 82.5)	−0.42 (−0.5, −0.3)	−0.57 (−0.7, −0.4)
High-missing, high-p					
0.7	40	0.3	56.57 (56, 57.1)	0.06 (−0.3, 0.4)	−0.08 (−0.4, 0.3)
0.7	40	1.0	68.5 (68, 69)	−0.03 (−0.3, 0.3)	−0.31 (−0.6, 0)
0.7	40	3.0	81.11 (80.8, 81.4)	−0.16 (−0.3, 0)	0.12 (0, 0.3)

random intercept for each generated sample and fixed effects for modeling type (where the lasso applied to the complete data is set to be the reference category). The resulting fixed effects can thus be interpreted as the expected change in the performance metric in terms of percentage points compared to the lasso applied to the complete (full) data set (i.e., without any missing observations). This particular model was chosen to account for the high amount of correlation in AUC, FDR, or FNR that arises due to the fact that the methods in question are fit to the same data set within each simulation. In other words, adding the random intercept accounts for the fact that some generated data sets are easier or harder to fit than others, and thereby allows more precision for the comparisons between the groups.

3.2 Simulation Results

The simulation results pertaining to predictive performance under the high correlation setting are shown in Figure 1, and their accuracy relative to the full-data lasso model is shown in Table 1. It is evident that MALCoM methods generally performed fairly similar to the MI-lasso procedure in terms of AUC, although in many of the simulated settings, at least one MALCoM method outperformed the MI-lasso. MALCoM methods performed comparably to the full-data lasso in terms of prediction; expected AUC was within 1 percentage point of the full-data AUC even when the missing proportion was 70%. MI-lasso, on the other hand, performed worse than the full-data AUC in low p settings and the high p, medium SNR setting. MALCoM-MCP was optimal in the high SNR region, but comparable or slightly worse than the other methods otherwise.

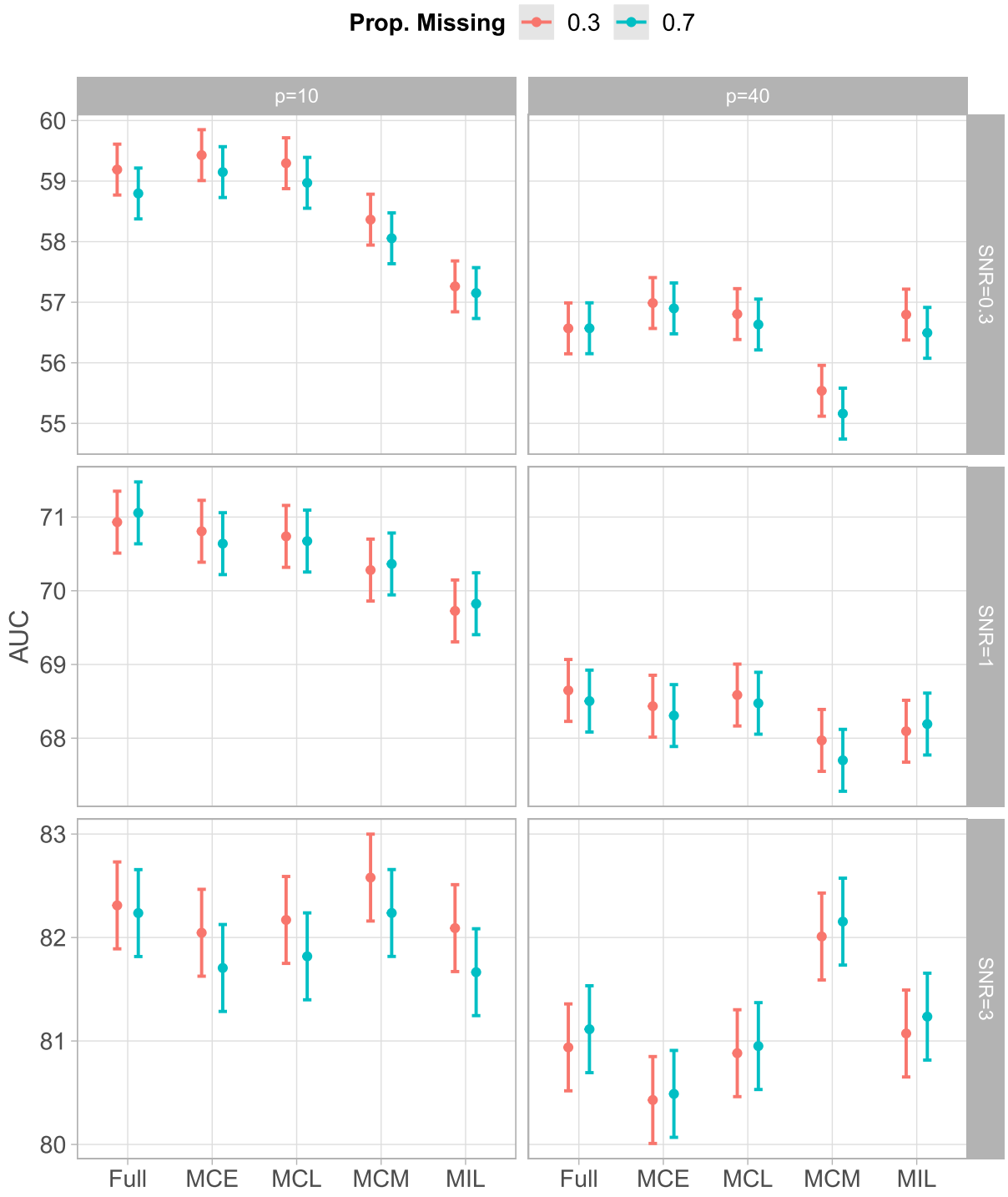

Figure 1: Estimated mean and 95% confidence intervals for predictive performance (area under ROC curve; AUC) for various methods under varied simulation settings with moderately high exchangeable feature correlation. MIL: Multiple imputation lasso; MCL: Median coefficient lasso; MCE: Median coefficient elastic net; MCM: Median coefficient minimax concave penalty. Raw values are presented via boxplots in Figure A1.

The MI-lasso generally had a lower false discovery rate and a higher false negative rate than the full data lasso and the MALCoM-lasso (Tables A1-A2 and Figures A2-A3). This difference in the weighting of false negatives vs positives is likely due at least in part to the difference in tuning method; BIC is generally more conservative than CV since the former attempts to select consistent models, while the latter attempts to minimize prediction error. However, MALCoM-MCP is able to more closely mirror the FDR and the FNR of the MI-lasso since MCP tends to selects sparser models, even when CV is used to tune its shrinkage parameter. Simulation results for the low-correlation AR(1) setting were similar and are shown in the appendix (Figures A4-A6 and Tables A3-A5).

To summarize, while the optimal method depends necessarily on the context and the true data generating mechanism (e.g. the signal to noise ratio, the number of candidate predictors, etc.), MALCoM is versatile enough to be implemented with the lasso, MCP, or elastic net penalties (among others), and therefore can thrive in situations where other competing methods struggle.

4 Application

4.1 Data and Methods

Data for this application consist of COVID-19 positive adult patients who were admitted to the University of Colorado Hospital in Aurora, Colorado between March 19, 2020 and April 2, 2020. The outcome of interest is whether the patient was intubated or not, and potential predictors of this were age, sex, body mass index (BMI), sex, race/ethnicity, presence of diabetes, symptom duration prior to admission, and certain biomarkers collected upon initial admission including lactate dehydrogenase (LDH), C-reactive protein (CRP), (logged) D-dimer, (logged) ferritin, and the neutrophil-to-lymphocyte ratio (NLR). Originally, 158 observations were collected and used for model training. Even though there were many fewer covariates than there were observations, penalized regression was deemed the best course of action because a) investigators desired a sparse predictive model, and b) there existed substantial correlation among covariates. After training the original MALCoM model, a second cohort of 102 patients consecutively admitted after April 2, 2020 with the same inclusion criteria, exclusion criteria, and chart data were followed prospectively in order to measure the performance of the model. This current application will follow a similar procedure for validation, where the MI-lasso and candidate alternative MALCoM models will be trained using the original 158 observations, and validated on the additional 102 observations. The data collection was reviewed and approved by the Colorado Multiple Institutional Review Board as exempt research.

No observations were missing the outcome, however 107 observations (67.7%) were missing at least one of the predictors. Those missing certain covariate data (e.g. LDH, ferritin, duration of symptoms before admission, and NLR) were found to differ by other features and by the outcome. Via the `mice` package in R, 20 augmented/imputed data sets were produced using chained equations to impute missing features sequentially. Each model in the imputation process consisted of a random forest with 50 trees.

Models were tuned as described in the previous section, and the optimal set of (aggregated) coefficients are obtained as described in Section 2. Each final set of coefficients was used to predict the intubation status for those in the prospective validation cohort, and an out-of-sample AUC, deviance-based R-squared, and misclassification rate were calculated to assess prediction accuracy. The deviance-based R-squared is defined as $R^2 = \frac{D_0 - D_1}{D_0}$, where

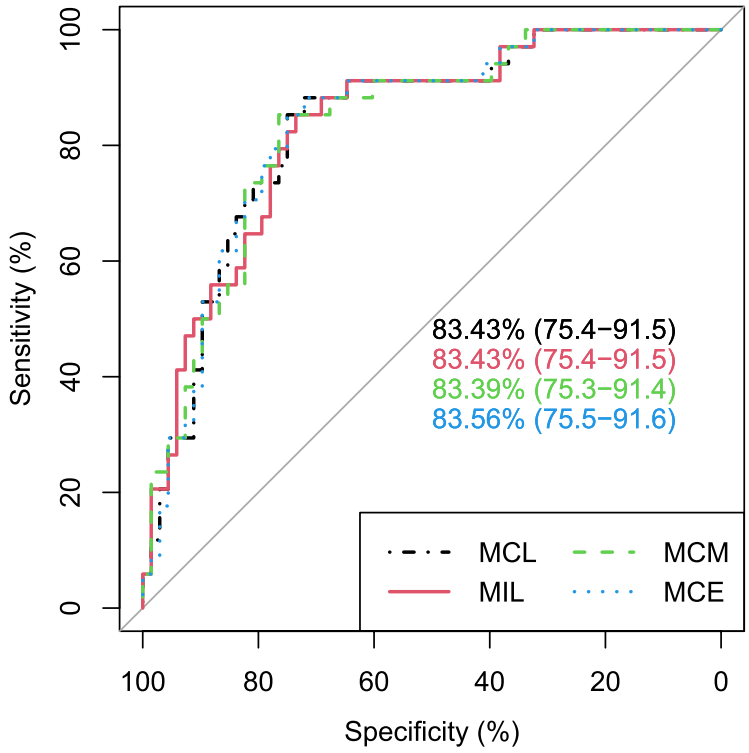

Figure 2: Receiver operating characteristic (ROC) curves for each model on test data set. MIL: Multiple imputation lasso; MCL: Median coefficient lasso; MCE: Median coefficient elastic net; MCM: Median coefficient minimax concave penalty; AUC: Area under ROC curve. AUC displayed with 95% bootstrapped confidence intervals (B = 2000).

$D_1 = -2\sum_i \{y_i \log \hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i)\}$, D_0 refers to the null model's deviance, and i indexes patients from the validation cohort.

If a patient in the prospective validation cohort had any missing data, mean imputation was performed prior to computing that patient's probability of intubation. However, the validation data for our application were almost fully complete; no variable was missing more than 10% of the observations, and the majority were missing less than 5%. Mean imputation was used for this small subset of patients because it is very simple to implement prospectively when using these predictive models in practice; since the continuous covariates were centered and scaled, performing mean imputation was as simple as plugging in a value of zero for the missing covariate's Z-score (and hence, not using that particular measure in the prediction equation). The only binary covariate in the final model was diabetes, which was not missing any values in the prospective validation set.

4.2 Results

Summary information about the population is shown in Table 2; the prospective cohort looked somewhat similar to the training cohort, with the exception of sex, ethnicity, coronary artery disease, and diabetes distribution. While ideally the validation set will look similar to the training set in most respects, differences indicate that the validation results are likely to be biased downward (i.e. models will appear less accurate than they truly are on the target population).

Table 2: Sample demographics. BMI: Body Mass Index; COPD: Chronic obstructive pulmonary disease; LDH: Lactic acid dehydrogenase; CRP: C-reactive protein; NLR: Neutrophil-lymphocyte ratio. Numbers represent counts (percentages) for categorical variables and means (standard deviations) for continuous variables.

	Training (n=158)	Validation (n=102)
Intubated	64 (40.5%)	34 (33.3%)
Age	56.15 (16.75)	55.09 (17.40)
BMI	31.34 (7.99)	31.28 (8.57)
Male	81 (51.3%)	67 (66.3%)
Tobacco Use	9 (5.8%)	3 (2.9%)
Ethnicity/Race		
White	30 (19.5%)	15 (14.7%)
Black	53 (34.4%)	17 (16.7%)
Hispanic	45 (29.2%)	46 (45.1%)
Asian	11 (7.1%)	14 (13.7%)
Other	15 (9.7%)	10 (9.8%)
Nonwhite	124 (80.5%)	87 (85.3%)
Diabetes	47 (30.3%)	47 (46.1%)
Hypertention	87 (56.1%)	60 (59.4%)
COPD	11 (7.1%)	8 (7.8%)
Cirrhosis	3 (1.9%)	0 (0.0%)
Coronary Artery Disease	18 (11.6%)	4 (3.9%)
Active Cancer	11 (7.1%)	3 (2.9%)
Immunosuppressed	10 (6.5%)	3 (2.9%)
Subjective Fever	111 (74.0%)	71 (71.7%)
Cough	126 (82.4%)	82 (82.8%)
Diarrhea	40 (26.8%)	32 (32.3%)
Nausea/Vomiting	42 (28.2%)	31 (31.3%)
Myalgias	34 (22.8%)	33 (33.3%)
Dyspnea	116 (75.3%)	73 (73.7%)
Duration of Symptoms	6.65 (4.62)	6.36 (4.58)
LDH	349.36 (135.55)	392.19 (253.09)
D-dimer	2468.98 (8991.93)	1412.32 (2847.03)
CRP	98.24 (78.34)	109.03 (81.98)
Ferritin	555.49 (822.41)	1186.88 (4014.01)
Neutrophils	5.96 (3.31)	5.90 (3.68)
Lymphocytes	2.26 (9.61)	1.19 (0.48)
NLR	6.40 (5.41)	6.53 (6.65)

Table 3 shows that the outcome of intubation is significantly related to missingness in several variables of interest, indicating that data are not missing completely at random. In particular, patients who were missing LDH, ferritin, duration of symptoms, and NLR were more likely to have been intubated. The optimal aggregated coefficients for each model are shown in Table 4, and their predictive performance is compared in Figure 2. Apparently, all of the modeling

Table 3: Number (percentage) missing for each variable of interest stratified by intubation status. BMI: Body Mass Index; LDH: Lactic acid dehydrogenase; CRP: C-reactive protein; NLR: Neutrophil-lymphocyte ratio.

	Not intubated (n=94)	Intubated (n=64)	Total (n=158)	p value
CRP	2 (2.1%)	6 (9.4%)	8 (5.1%)	0.062
LDH	5 (5.3%)	13 (20.3%)	18 (11.4%)	0.005
D-dimer	15 (16.0%)	14 (21.9%)	29 (18.4%)	0.404
Ferritin	7 (7.4%)	17 (26.6%)	24 (15.2%)	0.001
Duration of Symptoms	1 (1.1%)	6 (9.4%)	7 (4.4%)	0.018
Age	0 (0.0%)	1 (1.6%)	1 (0.6%)	0.405
BMI	5 (5.3%)	0 (0.0%)	5 (3.2%)	0.081
Nonwhite	1 (1.1%)	3 (4.7%)	4 (2.5%)	0.304
Diabetes	0 (0.0%)	3 (4.7%)	3 (1.9%)	0.065
Male	0 (0.0%)	0 (0.0%)	0 (0.0%)	
NLR	59 (62.8%)	24 (37.5%)	83 (52.5%)	0.002
Any of the above	65 (69.1%)	42 (65.6%)	107 (67.7%)	0.729

frameworks selected similar models, although the MCM model and the MI-lasso selected the fewest number of active coefficients. All models selected CRP, LDH, diabetes, and NLR as important features. The MCE and MCL models were more saturated, selecting indicators for non-white race and sex (though these coefficients were small in magnitude). The MCE model additionally selected log ferritin, though again the coefficient was small. The estimated nonzero coefficients varied considerably across imputations for all MALCoM methods (especially for MCE and MCL), indicating that analyses resulting from any single imputation procedure were sensitive to the random seed set prior to imputation (Tables A9-A11).

In terms of predictive accuracy, all four models performed similarly on the out-of-sample AUC (Figure 2). This pattern is also evident in Table 5, where the out-of-sample deviance and in the accuracy rate (i.e. the percent correctly classified using a 0.5 predicted probability cutoff) are also very similar across models, although the MALCoM-MCP performed slightly better on both measures. Due to its parsimony and its predictive performance, the MALCoM-MCP model was ultimately chosen as the final model for intubation risk. Given the fact that COVID-19 severity has been consistently linked to age and BMI in other studies, it seems odd that neither of these show up in as important in the final model. One likely reason which contributes to this is the fact that the patients under study are already quite severe, in that they are in the hospital for COVID-19-related reasons. The patients were generally older (mean age = 56) and had a high average BMI of 31.3. The prediction model answers the question: "given a patient has worse enough symptoms to come to a hospital for COVID-19, what characteristics are predictive of intubation risk?" It is highly likely that age and BMI are predictive of having to go to the hospital in the first place, but this is not addressed in the predictive model.

Table 4: Estimated aggregated coefficients from each modeling framework. MCL: Median coefficient lasso; MCE: Median coefficient elastic net; MCM: Median coefficient minimax concave penalty; BMI: Body Mass Index; LDH: Lactic acid dehydrogenase; CRP: C-reactive protein; NLR: Neutrophil-lymphocyte ratio. The column of MI-lasso refers to a change in estimated probability as opposed to log odds; for approximate MALCoM linear probability models, see Table A12 in Supplementary Material.

	MCE	MCM	MCL	MI-lasso*
Intercept	−0.63	−0.59	−0.56	0.39
CRP (Z-score)	0.66	0.92	0.69	0.13
LDH (Z-score)	0.31	0.27	0.30	0.04
Log D-dimer	-	-	-	-
Log ferritin	0.01	-	-	-
Symptom duration	-	-	-	-
Age	-	-	-	-
BMI	-	-	-	-
Nonwhite	0.06	-	0.00	-
Diabetes	0.48	0.58	0.49	0.06
Male	0.01	-	-	-
NLR	0.23	0.19	0.22	0.02

Table 5: Accuracy on test set for each modeling framework. MCL: Median coefficient lasso; MCE: Median coefficient elastic net; MCM: Median coefficient minimax concave penalty.

	AUC	R2	Accuracy
MCE	83.56	17.98	77.45
MCM	83.39	19.56	78.43
MCL	83.43	18.26	77.45
MI-lasso	83.43	15.94	76.47

5 Discussion

This work has focused on multiple imputation, but alternative strategies for handling missingness exist. Such methods include full-information maximum likelihood and inverse probability weighting, both of which are valid and often preferred (in theory) to list-wise deletion and single imputation. However, neither approach is feasible in this particular variable selection setting. Full-information maximum likelihood is less practical and versatile than multiple imputation methods, which can more easily account for auxiliary variables and have more readily usable software (Collins et al., 2001). Inverse probability weighting would work well if the missingness is predominantly in the outcome and the covariates are mostly complete, but it is less efficient in settings where covariates have many missing values since only the completely observed rows of the design matrix can be used to estimate the weights (Seaman and White, 2013). Therefore, multiple imputation seemed to be the optimal choice for this particular setting since it is flex-

ible enough to account for potentially many auxiliary variables and able to fully leverage the information contained in both the complete and incomplete variables.

Though initially a rather simple ad hoc method, I have shown empirically that MALCoM is competitive with other existing methods in terms of predictive accuracy and selection performance. Further, MALCoM's simplicity facilitates its use with existing software packages, such as those used for non-convex regularization or elastic net procedures. In the application, this versatility was valuable as it was able to produce a sparse model with satisfactory predictive performance. MALCoM requires each coefficient be selected with the same sign in a majority of imputations, whereas the MI-lasso selects each covariate across multiple imputations as a group. Though the MI-lasso ensures that selected variables across imputed data sets are the same and MALCoM does not, the aggregated coefficients after pooling yield a set of selected/deselected predictors in either setting. Theoretical properties of MALCoM methods under asymptotic conditions have not been studied.

The main benefit of MALCoM in comparison to its competitors is its ability to be used in settings where the lasso is known to be worse than other penalized regression approaches. When feature correlation is high, the elastic net method will perform comparably better than the lasso. When the signal-to-noise ratio is high, MCP will perform better than the lasso as well. So, pairing MALCoM with the right penalized regression framework for the application at hand will often yield optimal predictions and models. Further, in comparison to single-imputation procedures (e.g., Gong et al., 2020), MALCoM is less sensitive to the stochasticity resulting from imputation. In the application, the final MALCoM model was selected in only 5-35% of single imputed analyses (see Tables A9-A11). Therefore only performing the procedure once could result in a different predictive model due only to the random seed set prior to imputation. In conclusion, MALCoM is a convenient and versatile way to aggregate coefficients from penalized regression models fit to multiply imputed data sets.

Supplementary Material

A script to reproduce simulations under varied parameters has been provided as supplemental material online, along with an appendix containing additional tables and figures pertaining to the simulations described herein.

Acknowledgements

The author gratefully acknowledges support from Dr. Kristine Erlandson, Dr. Samuel Windham, and Dr. Kellen Hirsch, as well as the University of Colorado medical students who contributed to data collection. Dr. Erlandson coordinated and organized data collection and was the primary PI for the IRB protocol. Drs. Windham and Hirsch were instrumental in terms of developing the covariates of interest and shaping the prediction problem to be maximally clinically useful.

References

Breheny P, Huang J (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5(1): 232–253.

Chen Q, Wang S (2013). Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine*, 32(21): 3646–3659.

Collins L, Schafer JL, Kam C (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4): 330–351.

Friedman J, Hastie T, Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1): 1–22.

Gong J, Ou J, Qiu X, Jie Y, Chen Y, Yuan L, et al. (2020). A tool for early prediction of severe coronavirus disease 2019 (COVID-19): A multicenter study using the risk nomogram in Wuhan and Guangdong, China. *Clinical Infectious Diseases*, 71(15): 833–840.

Liu Y, Wang Y, Feng Y, Wall MM (2016). Variable selection and prediction with incomplete high-dimensional data. *The Annals of Applied Statistics*, 10(1): 418–450.

Long Q, Johnson BA (2015). Variable selection in the presence of missing data: Resampling and imputation. *Biostatistics*, 16(3): 596–610.

Meier L, Van De Geer S, Bühlmann P (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1): 53–71.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12: 77.

Rubin DB (2004). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Seaman SR, White IR (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3): 278–295. PMID: 21220355.

Sirimongkolkasem T, Drikvandi R (2019). On regularisation methods for analysis of high dimensional data. *Annals of Data Science*, 6(4): 737–763.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288.

Van Buuren S (2018). *Flexible Imputation of Missing Data*. CRC Press.

Van Buuren S, Groothuis-Oudshoorn K (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3): 1–67.

Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. (2020). Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ*, 369.

Yang Y, Yang H (2018). Model selection consistency of lasso for empirical data. *Chinese Annals of Mathematics, Series B*, 39(4): 607–620.

Zhang CH (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2): 894–942.

Zhao Y, Long Q (2017). Variable selection in the presence of missing data: Imputation-based methods. *WIREs Computational Statistics*, 9(5): e1402.

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