Heike Deutelmoser
Dr. sc. hum.

Development and Application of Robust Regularized Regression Methods in Multi-Omics Cancer Studies

Fach/Einrichtung: Medizinische Biometrie
Doktorvater: Prof. Dr. Justo Lorenzo Bermejo

Die vorliegende Arbeit beschäftigt sich mit dem Vergleich der regularisierten Regressionsmethode 'least absolute shrinkage and selection operator' (LASSO) mit zwei robusten Versionen von LASSO, und zwar dem robusten Huber-LASSO und dem robusten Quantil-LASSO. LASSO verhindert eine Überanpassung des Modells und erlaubt die Selektion der vielversprechendsten Teilmenge von Prädiktorvariablen. Standard-LASSO basiert jedoch auf der Verlustfunktion der kleinsten Quadrate, die bereits durch eine kleine Anzahl von Beobachtungen von Individuen mit abweichenden Phänotypen oder Genotypen stark beeinflusst werden kann, sogenannten Ausreißern oder 'high-leverage' Beobachtungen (Beobachtungen mit großem Einfluss). Robuste Verlustfunktionen können dieses Problem umgehen. Deren Potential in genetischen Studien molekularer Phänotypen wurde in dieser Arbeit untersucht.

Die alternativen LASSO-Methoden wurden unter Verwendung der folgenden Kriterien verglichen: Die Stabilität des Regularisierungsparameters, die Modell-Konsistenz, die Falsch-Positiv- und Richtig-Positiv-Rate und die Vorhersagegenauigkeit der Protein-, Metabolit- und messenger Ribonukleinsäure (mRNA) Expressionswerte. Die Analysen wurden auf der Grundlage von Genotyp- und molekularen Daten von Individuen durchgeführt. Der Regularisierungsparameter wurde durch zehnfache Kreuzvalidierung berechnet, und die Vorhersagegenauigkeit der molekularen Phänotypen wurde durch fünffache Kreuzvalidierung ausgewertet. Standard- und robuste Korrelationsmaße wurden untersucht, um die Vorhersagegenauigkeit zu beurteilen.

Ich führte umfassende Simulationen durch und wandte die untersuchten LASSO-Modelle auf reale Daten aus der INTERVAL-Studie, der Kooperative Gesundheitsforschung der Region Augsburg und dem 'Genotype-Tissue Expression'-Projekt an. Unterschiedliche Anteile simulierter Ausreißer wurden ausgewertet und die erklärte Varianz des Plasmaproteinspiegels, des Serummetabolitenspiegels und des mRNA-Expressionsspiegels wurde untersucht. Die statistische Tiefe und die erklärte Varianz wurden verwendet, um repräsentative unabhängige und abhängige Variablen für die Simulationsstudie und die realen Datenanwendungen auszuwählen.

Die Simulationsergebnisse zeigten, dass der robuste Huber-LASSO und der robuste Quantil-LASSO bessere Vorhersagen von Proteinwerten auf der Grundlage individueller Genotypdaten als der Standard-LASSO lieferten. Einzelne simulierte Ausreißer übten einen größeren Einfluss auf den Regularisierungsparameter, die Modell-Konsistenz und die Richtig-Positiv-Rate des Standard-LASSO aus als auf die des robusten Huber-LASSO und des robusten Quantil-LASSO. Die drei regularisierten Regressionsmethoden zeigten einen geringen Anteil an falsch-positiven Ergebnissen. Die Anwendungen der Methoden auf reale Daten bestätigten die Ergebnisse der Simulationsstudie.

Ich untersuchte die erklärte Varianz für drei häufige molekulare Phänotypen. Im Allgemeinen war die erklärte Varianz für die mRNA-Expression (1. Quartil = 0.0003, 3. Quartil = 0.0040)

geringer als die erklärte Varianz für Plasmaproteine (1. Quartil = 0.02, 3. Quartil = 0.09) und Serummetaboliten (1. Quartil = 0.04, 3. Quartil = 0.16). Für alle drei untersuchten molekularen Phänotype nahm die erklärte Varianz mit der Anzahl der assoziierten Einzelnukleotid-Polymorphismen zu.

Die drei Hauptneuheiten dieser Arbeit sind (1) die Untersuchung der Stabilität des Regularisierungsparameters unter Verwendung eines datenbasierten Ansatzes, (2) die Kombination von Ausreißern in den abhängigen und der unabhängigen Variable, und (3) die Untersuchung der Anzahl assoziierter genetischer Varianten für drei häufige molekulare Phänotypen. Zu den wichtigsten Einschränkungen gehörten die Abhängigkeit der Protein-, Metabolit- und mRNA-Expressions-Vorhersage von starken Assoziationen zwischen den genetischen Prädiktoren und dem molekularen Phänotyp sowie die Beschränkung der Auswertungen auf die regularisierte Regressionsmethode LASSO.

Die durchgeführten Vergleiche können als mögliche Leitfäden für zukünftige Forschung im Gebiet der robusten regularisierten Regressionsmethoden in genetischen Assoziationsstudien dienen. Andere robuste Versionen von regularisierten Regressionsmethoden, z.B. basierend auf 'Hampel's redescending' als auch die Robustifizierung populärer Vorhersageinstrumente wie PrediXcan sind das Ziel zukünftiger Forschung.