ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Transfer learning to minimize the predictive risk in clinical research

Samuel Branders, Jérôme Paul, Arthur Ooghe, and Alvaro Pereira

Cognivia
Rue Granbonpre 11, bte 9 1435 Mont-Saint-Guibert, Belgium

Abstract. The volume of data collected from patients enrolled in clinical trials is constantly on the rise. Classical linear and generalized linear models used in this context are unable to keep pace with this trend. Conversely, machine learning models have the potential to deal with such data, but cannot provide guarantees in terms of bias and interpretability. This paper explores a transfer learning approach that seeks to harmonize the strengths of both paradigms: providing unbiased and interpretable estimators while minimizing the expected predictive risk in finite samples.

1 Introduction

The volume of data collected from the limited number of patients enrolled in clinical trials is constantly on the rise. Accessing genomic, proteomic, and activity data has never been more convenient. In the context of clinical research, the use of these data holds the promise of crucial insights into patient characterization and a deeper comprehension of treatment efficacy. However, the magnitude and diversity of these datasets present important challenges for their analysis.

Analyses of clinical trials often aim to estimate the effect of a few parameters of interest, such as the treatment and its interaction with age, and sex, on the evolution of a disease. Traditional statistical methods, such as ANOVA, ANCOVA, logistic regression, or Cox models, are commonly used because of their good statistical properties and high interpretability. Yet, they fall short when trying to improve the estimation of these parameters of interest while adjusting for more complex data *e.g.* featuring an extensive number of variables or more complex structures. Conversely, more complex machine learning models deal with these data but often at the cost of reduced interpretability or the introduction of biases due to their assumptions *e.g.* in ridge regression or lasso [1].

This paper presents a transfer learning approach that seeks to harmonize the strengths of both paradigms: the performances and adaptability of machine learning with the assurances of interpretability and unbiasedness of classical statistical methods.

2 Generative model of the data

We are here looking to estimate precisely a set of chosen parameters in the presence of noise. The response model writes

$$Y = \mu + \gamma^T Z + U \tag{1}$$

where $Z = (Z_1, \ldots, Z_d)^T$ is a vector of d variables of interest. γ is the vector of parameters we are looking to estimate, U is the error term, and μ is the

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

intercept. The random variable $U \sim \mathcal{N}(0, \sigma_u^2)$ accounts for factors not linked with treatment and is assumed to be independent of Z. This independence is guaranteed in randomized clinical trials where Z represents the treatment assignment or interactions with the treatment. The random variable U may in turn be expressed as a function, φ, of a vector of p covariates, $X = (X_1, \ldots, X_p)^T$ and an error term.

$$U = \varphi(X) + \varepsilon \tag{2}$$

The error term ε is assumed to be normally distributed $\mathcal{N}(0, \sigma_\varepsilon^2)$, independently of X and Z. We have divided our model into two parts, Z and X, to highlight the role of each part. Z are the variables of interest and X are covariates only used for improving the estimation of γ.

The first part of this paper assumes a linear dependence between the covariates and the response: $\varphi(X) = \beta^T X$. Thus, by combining Equations (1) and (2) and assuming $\mu = 0$ without loss of generality, the general regression model writes

$$Y = \gamma^T Z + \beta^T X + \varepsilon. \tag{3}$$

The non-linear $\varphi(X)$ is discussed at the end of this paper.

Now, consider a random sample of n observations where $\boldsymbol{Z} \in \mathbb{R}^{n \times d}$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ represent respectively the matrices of observed patients' variables and covariates. Without loss of generality, we can assume that the columns of $\boldsymbol{Z}$ are standardized and orthogonal to each other. Hence, the gram matrix of $\boldsymbol{Z}$ is diagonal, $\boldsymbol{Z}^T\boldsymbol{Z} = (n-1)I_d$. Indeed, this paper aims to discuss the benefit of $\boldsymbol{X}$ in the estimation of γ irrespective of the collinearity between the variables of interest.

3 Predictive risk

To evaluate the relevance of the information contained in $\boldsymbol{X}$ for the evaluation of γ, we can compare the evolution of the predictive risk using the same framework as in [1]. Consider a test point z_0, independent of the training data. For the estimator $\hat{\gamma}$, its out-of-sample prediction risk (or 'risk') can be defined as:

$$R_{Z,X}(\hat{\gamma}, \gamma) = \mathbb{E}[(z_0^T\hat{\gamma} - z_0^T\gamma)^2 | \boldsymbol{Z}, \boldsymbol{X}] = \mathbb{E}[\|\hat{\gamma} - \gamma\|_{\Sigma_z}^2 | \boldsymbol{Z}, \boldsymbol{X}]$$

where $\|x\|_{\Sigma_z}^2 = x^T\Sigma_z x$, and Σ_z is the covariance matrix of $\boldsymbol{Z}$. This definition of the risk is conditional on the variable of interest Z (the target for the estimation) and all other variables for the adjustment X (only used to have a more precise estimation of γ). Of note, our risk definition slightly differs from [1] as it does not include $\hat{\beta}$. Indeed, we consider here that only γ is of interest and not β.

This risk can be decomposed between bias and variance:

$$R_{Z,X}(\hat{\gamma}, \gamma) = \underbrace{\|\mathbb{E}(\hat{\gamma}|\boldsymbol{Z}, \boldsymbol{X}) - \gamma\|_{\Sigma_z}^2}_{B_{Z,X}(\hat{\gamma},\gamma)} + \underbrace{Tr[\mathrm{Var}(\hat{\gamma}|\boldsymbol{Z}, \boldsymbol{X})\Sigma_z]}_{V_{Z,X}(\hat{\gamma},\gamma)}$$

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

4 Predictive risk of using covariates

To estimate the added value of the covariates $\boldsymbol{X}$, we should compare the predictive risk of the estimators of γ with and without them. To avoid any confusion, we denote by $\hat{\gamma}_0$ the ordinary least squares (OLS) estimator of γ when no covariate is used in the regression (Equation 1) and by $\hat{\gamma}_p$ when p covariates are included in the regression (Equation 3).

The variance-covariance matrix of $\hat{\gamma}_0$ conditional on $\boldsymbol{Z}$ is:

$$\mathrm{Var}(\hat{\gamma}_0|\boldsymbol{Z}) = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\sigma_u^2 = \boldsymbol{D}\sigma_u^2 \tag{4}$$

where $\boldsymbol{D}$ is a diagonal matrix whose diagonal elements are $\boldsymbol{D}_{kk} = \frac{1}{n-1}$. The variance-covariance matrix of $\hat{\gamma}_p$ conditional on $\boldsymbol{X}$ and $\boldsymbol{Z}$ is:

$$\mathrm{Var}(\hat{\gamma}_p|\boldsymbol{Z},\boldsymbol{X}) = \left[\boldsymbol{D} + \boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{X}S^{-1}\boldsymbol{X}^T\boldsymbol{Z}\boldsymbol{D}\right]\sigma_\varepsilon^2 \tag{5}$$

where $S = \boldsymbol{X}^T(I_n - \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T)\boldsymbol{X}$. Ordinary least squares estimators are unbiased [1]. Knowing that Σ_z is I_d per construction, the risk can be computed from the variance-covariance matrix:

$$R_{Z,X}(\hat{\gamma}_p,\gamma) = Tr[\mathrm{Var}(\hat{\gamma}_p|\boldsymbol{Z},\boldsymbol{X})\Sigma_z] = \sigma_\varepsilon^2\sum_{k=1}^{d}\left[\boldsymbol{D} + \boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{X}S^{-1}\boldsymbol{X}^T\boldsymbol{Z}\boldsymbol{D}\right]_{k,k} \tag{6}$$

The risk is conditioned on $\boldsymbol{Z}$ and $\boldsymbol{X}$. $\boldsymbol{X}$ is not known and should thus be treated as a random variable. A common assumption is to assume that the covariates are isotropic and generated with multivariate normal distribution $X \sim N(0,\Sigma)$ where $\Sigma = \boldsymbol{I}_p$. Then, we can compute the expectation of $R_{Z,X}$ from Equation 6 and using the linearity of the expectation:

$$\mathbb{E}_{Z,X}[R_{Z,X}(\hat{\gamma}_p,\gamma)] = \sigma_\varepsilon^2\sum_{k=1}^{d}E_{Z,X}\left[\left[\boldsymbol{D} + \boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{X}S^{-1}\boldsymbol{X}^T\boldsymbol{Z}\boldsymbol{D}\right]_{k,k}\right]$$

From papers [2, 3], we know that:

$$\frac{\left[\boldsymbol{D} + \boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{X}S^{-1}\boldsymbol{X}^T\boldsymbol{Z}\boldsymbol{D}\right]_{k,k}}{\boldsymbol{D}_{kk}} = \frac{1}{1-B}$$

where $B \sim \mathrm{Beta}(p/2,(n-d-p+1)/2)$. As such, the expected risk becomes:

$$\mathbb{E}_{Z,X}[R_{Z,X}(\hat{\gamma}_p,\gamma)] = \sigma_\varepsilon^2\frac{d}{n-1}\frac{(n-d-1)}{(n-d-p-1)} \tag{7}$$

The expected risk of the model without covariate is much simpler to compute and is simply equal to $\sigma_u^2 d/(n-1)$. The covariates are expected to be useful if the expected risk is lower when using them:

$$\begin{aligned}
\mathbb{E}_{Z,X}[R_{Z,X}(\hat{\gamma}_p,\gamma)] &< \mathbb{E}_Z[R_Z(\hat{\gamma}_0,\gamma)] \\
\sigma_\varepsilon^2\frac{d}{n-1}\frac{(n-d-1)}{(n-d-p-1)} &< \sigma_u^2\frac{d}{n-1} \\
\frac{p}{n-d-1} &< 1 - \frac{\sigma_\varepsilon^2}{\sigma_u^2} = \frac{\mathrm{SNR}}{\mathrm{SNR}+1}
\end{aligned} \tag{8}$$

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Where $SNR = (\sigma_u^2 - \sigma_\varepsilon^2)/\sigma_\varepsilon^2$ is the signal-to-noise ratio of the covariates and $1 - \sigma_\varepsilon^2/\sigma_u^2$ is the proportion of the variance of U explained by the covariates. This inequality shows that the benefit of using covariates is a trade-off between degrees of freedom (the number p of covariates) and their variance explained.

This result is consistent with the limiting result in [1] showing that the risk of standard linear regression is lower than the risk of the null model if $p/n < SNR/(SNR+1)$ when $n, p \to \infty$ and $p/n < 1$. However, the current result is particularly interesting in a clinical setting. Clinical studies often have limited numbers of patients for ethical reasons and our results show that these limiting results still apply in this context.

In a real clinical setting, the prognostic of patients is explained by numerous factors and most of these factors will have a very low SNR [4]. From Equation 8, numerous covariates and low SNR are not compatible. Adjusting for a large number of covariates, p, is only possible if the SNR is also large. This is why it is often recommended to limit the number of covariates in clinical trials.

As a solution, J. Tukey proposed 30 years ago the use of composite covariates [4, 5]. The idea is to combine several covariates (while guessing their importance) into one composite covariate to improve the adjustment without paying the cost in degrees of freedom. However, this approach has not been used in practice and its effectiveness is difficult to assess. The objective of the next section is to propose a new perspective on this old idea using machine learning and the recent random matrix theory results to estimate the predictive risk of the estimators. In particular, this old idea can be framed as transfer learning.

5 Transfer learning

Transfer learning is commonly described as a method where a model developed for one task is reused to improve the performance of a model on another related task. Here, we propose to learn a model $\hat{\varphi}(X) = \hat{\beta}^T X$ using previous independent data and to use the model's prediction $\hat{W} = \hat{\varphi}(\boldsymbol{X})$ as a covariate to improve the estimation of γ. Using $\hat{W}$ allows to transfer the knowledge extracted from independent data to avoid having to learn both β and γ on the same study.

$$Y = \gamma Z + \alpha \hat{W} + \varepsilon_w$$

The covariates X are usually known factors influencing patient prognosis. They are not specific to the treatment or condition studied. Thus, $\hat{\varphi}(X)$ is a prognostic model and $\hat{W}$ the patient's prognostic score. While adjusting for patient prognostic is relevant, reestimating β in small studies is not of primary interest. Therefore, we intend to leverage available historical data of possibly a much larger sample size m than n the number of patients available to estimate γ.

As $\hat{W}$ is estimated, the error term ε_w now accounts for both the prediction error on $\hat{\varphi}$ and the error term of the initial model, ε. As $\hat{\varphi}$ is estimated from independent data, the two parts of ε_w are independent. We thus have:

$$\sigma_{\varepsilon_w}^2 = \alpha R_X(\hat{\beta}, \beta) + \sigma_\varepsilon^2$$

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

where $R_X(\hat{\beta}, \beta)$ is the predictive risk of the prognostic model trained on independent data. The weight of $\hat{W}$ in the model is denoted α and has an expected value of $\text{Var}(\beta^T X)/(\text{Var}(\beta^T X)+R_X(\hat{\beta}, \beta))$. Then, it is easy to show that $\sigma_\varepsilon^2 \leq \sigma_{\varepsilon_w}^2 \leq \sigma_u^2$. These lower and upper bounds are respectively reached when the predictive risk on $\hat{\beta}$ is 0 and ∞. When $R_X(\hat{\beta}, \beta) \to 0$ the prognostic model is perfect $\alpha \to 1$ and $\sigma_{\varepsilon_w}^2 \to \sigma_\varepsilon^2$. When the risk becomes large, the usefulness of $\hat{W}$ is low, $\alpha \to 0$ and $\sigma_{\varepsilon_w}^2 \to \sigma_u^2$. The expected risk of $\hat{\gamma}_w$ can be computed using the Equation 7 when there is only one covariate ($\hat{W}$):

$$\mathbb{E}_{Z,X}[R_{Z,X}(\hat{\gamma}_w, \gamma)] = \sigma_{\varepsilon_w}^2 \frac{d}{n-1} \frac{(n-d-1)}{(n-d-2)} \leq \mathbb{E}_Z[R_Z(\hat{\gamma}_0, \gamma)] \frac{(n-d-1)}{(n-d-2)}$$

As we can see, the expected risk of $\hat{\gamma}_w$ is not directly impacted by the number of covariates, p. With the bounds $\sigma_{\varepsilon_w}^2$, we can conclude that the use of a prognostic model is safe in terms of risk. Indeed, even if the prognostic model dramatically fails, the expected risk remains bounded and only marginally increases by a $(n-d-1)/(n-d-2)$ factor as compared to the model without covariate. This multiplicative factor rapidly becomes negligible when n increases. In contrast, the risk of $\hat{\gamma}_p$ is not bounded and depends on p the number of covariates.

Another advantage of this formulation is to decouple the estimation of γ and β. The estimation of γ will remain unbiased as long as $\hat{\varphi}$ is estimated from independent data. As such, we can use any machine learning approach. In particular, we could select an approach minimizing the predictive risk, $R_X(\hat{\beta}, \beta)$, regardless of its bias. For example, we could use a ridge regression which has a lower predictive risk than a standard linear regression [1]. In practice, this risk $R_X(\hat{\beta}, \beta)$ could be estimated using cross-validation or external data. Or, we could approximate it with the limiting risk when $n \to \infty$ using the random matrix theory [6, 7, 1]. $R_X(\hat{\beta}, \beta)$ can then be used to compute the expected risk of $\hat{\gamma}_w$ when using the transfer learning from historical data. This expected risk will be lower than the classical approach if $\sigma_{\varepsilon_w}^2 < \sigma_\varepsilon^2(n-d-p-1)/(n-d-2)$.

It should be noted that many machine learning algorithms are proven to be consistent and thus converge in probability to the true value ([8, 9, 6]). This means that the risk, $R_X(\hat{\beta}, \beta)$, will converge in probability to 0 if m the number of patients used to fit the prognostic model tends to infinity with $p/m \to 0$. As such, if there is enough historical data, the lower bound on the expected risk of $\hat{\gamma}_w$ can be reached and we have:

$$\mathbb{E}_{Z,X}[R_{Z,X}(\hat{\gamma}_w, \gamma)] = \mathbb{E}_{Z,X}[R_{Z,X}(\hat{\gamma}_p, \gamma)] \frac{(n-d-p-1)}{(n-d-2)}$$

The expected risk when using a prognostic model will be better than the model with the p covariates by a factor $(n-d-p-1)/(n-d-2)$. When p is large or n is small, the benefit of the transfer learning will be large and could be performed even if $\hat{\varphi}$ is not perfect or trained on a large dataset. In some simulations, the transfer learning was better even when $n \approx m$.

Of note, this section was written assuming the covariates to be linear but it is not strictly needed. One can trivially replace $\hat{W} = \hat{\beta}^T \boldsymbol{X}$ by $\hat{W} = \hat{\varphi}(\boldsymbol{X})$.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

The only difference is that the initial linear model with p covariates will not be able to correctly adjust for the non-linearities, further increasing the advantage of using a non-linear prognostic model trained from independent data.

6 Discussion and conlusion

In this paper, we propose a new perspective on an old problem: how to optimize the use of covariates/data to estimate some parameters of interest when the number of samples, n, is limited. This problem is often encountered in clinical trials when estimating the treatment effect from a limited number of patients while using many prognostic variables.

The proposed approach uses external data to fit a prognostic model using machine learning. The prognostic model predictions can then be used in the adjustment to improve the estimation of the parameters of interest. Assuming isotropic normally distributed covariates, we were able to compute the expected predictive risk of using the covariates or a prognostic model highlighting the similarity with limiting results (when $n, p \to \infty$) built on the recent advance in random matrix theory [1]. In particular, when enough external data are available, we showed that the use of a prognostic model can greatly improve the predictive risk. Furthermore, the estimation of the parameters of interest will remain unbiased and the predictive risk will only be marginally increased even if the prognostic model dramatically fails to predict anything.

References

[1] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

[2] Gwowen Shieh. Power Analysis and Sample Size Planning in ANCOVA Designs. *Psychometrika*, 85(1):101–120, 2020.

[3] Samuel Branders, Alvaro Pereira, Guillaume Bernard, Marie Ernst, Jamie Dananberg, and Adelin Albert. Leveraging historical data to optimize the number of covariates and their explained variance in the analysis of randomized clinical trials. *Statistical Methods in Medical Research*, 31(2):240–252, 2022.

[4] John W Tukey. Use of Many Covariates in Clinical Trials. *International Statistical Review / Revue Internationale de Statistique*, 59(2):123–137, 1991.

[5] John W. Tukey. Tightening the clinical trial. *Controlled Clinical Trials*, 14(4):266–285, 1993.

[6] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *Annals of Statistics*, 46(1):247–279, 2018.

[7] Bingxin Zhao, Fei Zou, and Hongtu Zhu. Cross-Trait Prediction Accuracy of Summary Statistics in Genome-Wide Association Studies. *Biometrics*, 79(2):841–853, jun 2023.

[8] Erwan Scornet, Gerard Biau, and Jean Philippe Vert. Consistency of random forests. *Annals of Statistics*, 43(4):1716–1741, 2015.

[9] Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep Neural Networks for Estimation and Inference. *Econometrica*, 89(1):181–213, 2021.