Risk Prediction for Partially Heterogeneous Subgroups via Fusion

Eun Jeong Oh [1] Runze Li [2] Ravi B. Parikh [3] Jinbo Chen [1]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA
[2]Department of Statistics, Penn State University, University Park, PA
[3]Department of Medical Ethics and Health Policy, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Motivation

Accurate risk modeling is challenging due to variation in baseline risk and risk predictors across patient subgroups. Such heterogeneity in risk, if left unrecognized, can lead to compromised accuracy. The data for subgroups may not be sufficiently rich to allow separate analyses when the number of predictors is large. To overcome this, we propose a novel algorithm to fit subgroup-specific models, which leverages the sharing of a common predictor while performing variable selection for subgroup-specific predictors. Building upon an existing fusion technique, the proposed method encourages similarity among subgroup-specific parameters for the common predictor.

A Partially Heterogeneous Subgroup-Specific Risk Prediction Model

- Y: a binary outcome; $\mathbf{X}$: a p-dimensional EHR predictors
- S: a predictor that is highly predictive of Y (e.g., a predictive score from an existing model)
- K: the number of patient subgroups, which is defined *a priori*
- $\mathcal{G} \equiv \{\mathcal{G}_1 \ldots, \mathcal{G}_K\}$: a partition of the data, with $|\mathcal{G}_k| = n_k$ for $k = 1, \ldots, K$ and $\sum_k n_k = N$

For each $i \in \mathcal{G}_k$, we consider $\text{logit}P(Y_i = 1|S_i, \mathbf{X}_i) = \alpha_k + S_i\beta_k + \mathbf{X}_i^\mathsf{T}\tau_k$, $k = 1, \ldots K$. To recognize that S may calibrate well in some of the K groups, we propose a fusion technique on $\{\alpha_k\}_{k=1}^K$ and $\{\beta_k\}_{k=1}^K$ while selecting subgroup-specific predictors through $\{\tau_k\}_{k=1}^K$, $\tau_k \in \mathbb{R}^p$, $k = 1, \ldots, K$.

Denote $\mathbf{\Phi}_i^\mathsf{T}\theta_k \equiv (1, S_i, \mathbf{X}_i)^\mathsf{T}(\alpha_k, \beta_k, \tau_k^\mathsf{T})^\mathsf{T}$, and let $\theta \equiv \{\theta_k, k = 1, \ldots, K\}$. We then fit the model by minimizing the penalized negative $\frac{1}{N}\sum_{k=1}^K \ell_k(\theta_k) + \rho_\lambda(\theta)$, log-likelihood function, where $\ell_k(\theta_k) = \sum_{i\in\mathcal{G}_k}[\psi(\mathbf{\Phi}_i^\mathsf{T}\theta_k) - Y_i\mathbf{\Phi}_i^\mathsf{T}\theta_k]$ is the kth subgroup-specific likelihood, $\psi(t) = \log(1 + \exp(t))$, and

$$\rho_\lambda(\theta) = \lambda\big(\sum_{k=2}^K |\alpha_{(k)} - \alpha_{(k-1)}| + c_\alpha|\alpha_{[1]}| + \sum_{k=2}^K |\beta_{(k)} - \beta_{(k-1)}| + c_\beta|\beta_{[1]}| + \sum_{k=1}^K J(\tau_k)\big).$$

- $\alpha_{(k)}$, $\beta_{(k)}$: the kth smallest component of $(\alpha_1, \ldots, \alpha_K)$ and $(\beta_1, \ldots, \beta_K)$, respectively
- $\alpha_{[1]}$, $\beta_{[1]}$: the smallest element of vector $(|\alpha_1|, \ldots, |\alpha_K|)$ and $(|\beta_1|, \ldots, |\beta_K|)$, respectively
- $c_\alpha \in \{0, 1\}$, $c_\beta \in \{0, 1\}$: a pre-determined constant for turning on or off sparsity

An Iterative Algorithm

We illustrate an iterative algorithm under $J(\tau_k) = \|\tau_k\|_1$, $k = 1, \ldots, K$ as follows.

- Initialization. For $k = 1, \ldots, K$, obtain initial subgroup-specific estimators

$$(\hat\alpha_k, \hat\beta_k, \hat\tau_k) \leftarrow \underset{\alpha_k,\beta_k,\tau_k}{\arg\min}\frac{1}{n_k}\ell_k(\alpha_k, \beta_k, \tau_k) + \lambda\sum_{j=1}^p |\tau_{kj}|.$$

- Step 1. For $k = 1, \ldots, K$, update $\hat\tau_k \leftarrow \arg\min_{\tau_k}\frac{1}{n_k}\ell_k(\hat\alpha_k, \hat\beta_k, \tau_k) + \lambda\sum_{j=1}^p |\tau_{kj}|$.
- Step 2.
  - Update

$$\hat{\boldsymbol{\alpha}} \leftarrow \underset{\alpha}{\arg\min}\frac{1}{N}\sum_{k=1}^K \ell_k(\alpha_k, \hat\beta_k, \hat\tau_k) + \lambda\Big(\sum_{k=2}^K |\alpha_{(k)} - \alpha_{(k-1)}| + c_\alpha|\alpha_{[1]}|\Big).$$

  - Update

$$\hat{\boldsymbol{\beta}} \leftarrow \underset{\beta}{\arg\min}\frac{1}{N}\sum_{k=1}^K \ell_k(\hat\alpha_k, \beta_k, \hat\tau_k) + \lambda\Big(\sum_{k=2}^K |\beta_{(k)} - \beta_{(k-1)}| + c_\beta|\beta_{[1]}|\Big).$$

- Repeat Step 1-2 until convergence.

The Local Optimality Property of Stationary Points (Bertsekas, 1999)

We derive the upper bound on the squared ℓ_2-error for the distance between stationary points (Bertsekas, 1999) and the corresponding population optimum in the theorem below. The theorem implies that the above coordinate descent algorithm is guaranteed to converge to stationary points within close proximity of the true parameter values.

Theorem 1. *Suppose (λ, R) are chosen such that α^*, β^*, and $\tau^* = (\tau_1^{*\mathsf{T}}, \ldots, \tau_K^{*\mathsf{T}})^\mathsf{T}$ are feasible and*

$$c\sqrt{\frac{\log((p+2)K)}{N}} \le \lambda \le \frac{c'}{R},$$

where (c, c') are some positive constants. Assume regularity conditions (A1)–(A4) in the Appendix hold. Then for any $N \ge CR^2\log((p+2)K)$ with a sufficiently large constant $C > 0$, with the probability at least $1 - c_1\exp(-c_2\log((p+2)K))$, any stationary points, $\tilde\alpha$, $\tilde\beta$ and $\tilde\tau$ of the objective function satisfy the estimation error bound

$$\left\|\begin{pmatrix}\tilde\alpha - \alpha^* \\ \tilde\beta - \beta^* \\ \tilde\tau - \tau^*\end{pmatrix}\right\|_2^2 \le \frac{c_0^2\lambda^2\kappa}{(4\gamma_1 - 3\mu)^2 b},$$

where $\kappa = \||(\alpha^{*\mathsf{T}}, \beta^{*\mathsf{T}}, \tau^{*\mathsf{T}})^\mathsf{T}\|_0$, c_0 is a positive constant, γ_1 is a constant depending on $\|(\alpha^{*\mathsf{T}}, \beta^{*\mathsf{T}}, \tau^{*\mathsf{T}})^\mathsf{T}\|_2$, ψ, $\lambda_{min}(\mathbf{\Sigma})$, and the sub-Gaussian parameters. Here, we assume $\mu < 2\gamma_1$.

Simulation Studies

- $\mathbf{X}_i \sim N(0, \Omega^k)$ and S_i obtained by plugging $\mathbf{X}_i$ into a preliminary risk model
- $Y_i \sim Bern(\pi_i)$ with $\pi_i = \exp(\eta_i)/[1 + \exp(\eta_i)]$, where $\eta_i = \alpha_k + S_i\beta_k + \mathbf{X}_i^\mathsf{T}\tau_k$
- $K = 10$, $p = 50$, and equal subgroup sizes $n_1 = \cdots = n_K = n$ varying from 100 to 250
- To introduce heterogeneity into α and β, we set $m = \lceil(1 - r)K\rceil$ elements of α and β to be homogeneous where $r \in \{0.3, 0.7\}$. For a given r value, we consider

$$\alpha_1 = \cdots = \alpha_m = -1.5, \qquad \beta_1 = \cdots = \beta_m = 1.5,$$
$$\alpha_j = -0.2 - 0.1j \quad \text{for } m < j \le K; \qquad \beta_j = 1.9 + 0.1j \quad \text{for } m < j \le K.$$

- $\tau_k = \{\mathbf{0}_{p(k-1)/K}, \text{seq}(1, 0.2, p/K), \mathbf{0}_{p(K-k)/K}\}$ if $1 \le k \le \lceil K/2\rceil$, and $\tau_k = \mathbf{0}_p$ otherwise

The proposed fused lasso approach ("FL") is compared to subgroup-specific analysis with lasso ("SL") and analysis of data aggregated from all subgroups with lasso ("CL")

Table 1. Simulation results on mean AUPRC, median calibration slope, MEDSE of the regression ceofficient for S, and mean Spearman's rank correlation ρ based on 1000 replications.

		AUPRC			Calibration Slope			MEDSE			Spearman's ρ		
r	Group	SL	CL	FL	SL	CL	FL	SL	CL	FL	SL	CL	FL
0.3	$\mathcal{G}_1$	0.75	0.53	0.76	0.85	1.10	0.83	2.25	0.17	0.46	0.84	0.42	0.84
	$\mathcal{G}_2$	0.72	0.50	0.72	0.85	1.08	0.83	2.25	0.17	0.56	0.86	0.30	0.85
	$\mathcal{G}_3$	0.60	0.54	0.65	0.80	1.42	0.77	2.25	0.17	0.56	0.74	0.50	0.74
	$\mathcal{G}_4$	0.76	0.57	0.77	0.87	1.61	0.85	2.25	0.17	0.56	0.89	0.52	0.90
	$\mathcal{G}_5$	0.65	0.45	0.68	0.82	0.86	0.81	2.25	0.17	0.56	0.81	0.29	0.83
	$\mathcal{G}_6$	0.02	0.40	0.39	-0.03	0.07	0.88	2.25	0.17	0.56	0.01	0.45	0.72
	$\mathcal{G}_7$	0.01	0.28	0.29	-0.06	0.02	0.87	2.25	0.17	0.56	0.00	0.25	0.74
	$\mathcal{G}_8$	0.03	0.67	0.69	0.03	0.08	0.89	7.29	1.82	3.80	0.01	0.38	0.74
	$\mathcal{G}_9$	0.04	0.62	0.67	0.04	0.12	0.94	7.84	2.10	4.12	0.02	0.42	0.89
	$\mathcal{G}_{10}$	0.06	0.60	0.62	-0.02	0.11	0.94	8.41	2.40	4.58	0.02	0.39	0.87
0.7	$\mathcal{G}_1$	0.75	0.53	0.75	0.86	1.20	0.83	2.25	1.08	0.59	0.88	0.48	0.82
	$\mathcal{G}_2$	0.72	0.51	0.73	0.85	1.16	0.84	2.25	1.08	1.27	0.88	0.35	0.86
	$\mathcal{G}_3$	0.60	0.55	0.64	0.79	1.48	0.78	2.25	1.08	1.27	0.83	0.59	0.75
	$\mathcal{G}_4$	0.85	0.74	0.86	0.88	1.77	0.86	5.29	3.38	3.71	0.91	0.63	0.90
	$\mathcal{G}_5$	0.80	0.62	0.82	0.84	0.99	0.83	5.76	3.76	3.50	0.87	0.34	0.86
	$\mathcal{G}_6$	0.02	0.68	0.70	-0.04	0.08	0.91	6.25	4.16	3.76	0.24	0.23	0.70
	$\mathcal{G}_7$	0.04	0.50	0.51	-0.08	0.00	0.96	6.76	4.57	4.08	0.13	0.00	0.72
	$\mathcal{G}_8$	0.03	0.67	0.69	0.03	0.05	1.07	7.29	5.01	4.53	0.23	0.14	0.75
	$\mathcal{G}_9$	0.04	0.60	0.66	-0.04	0.09	1.12	7.84	5.47	4.93	0.25	0.20	0.75
	$\mathcal{G}_{10}$	0.06	0.59	0.61	-0.08	0.06	1.07	8.41	5.95	5.71	0.23	0.13	0.72

Analysis of Penn Medicine EHR Data

Goal: predict 180-day risk of mortality for oncology patients using the structured data extracted from the University of Pennsylvania Health System EHRs

- $N = 20,723$ patients, ranging from 330 with thyroid cancer to 4,665 with breast cancer
- $K = 11$ cancer types
- $p = 198$ EHR predictors (e.g., lab results, comorbidities, demographics)
- An existing gradient boosting model (Parikh et al., 2019), which did not distinguish cancer types, was used to generate risk predictor S

Date split into two subsets of equal sizes that are used as the training and test sets, respectively.

Table 2. Estimated AUPRC and calibration slope for models of short-term mortality risk

	AUPRC			Calibration Slope		
Cancer Type	SL	CL	FL	SL	CL	FL
Breast	0.44	0.46	0.44	1.08	1.48	1.08
GI	0.43	0.44	0.43	1.15	0.80	1.14
GU	0.48	0.50	0.48	1.17	1.26	1.13
Gyn	0.42	0.43	0.42	0.96	1.15	0.77
Leukemia	0.33	0.32	0.33	0.83	0.81	0.79
Lymphoma	0.32	0.30	0.33	1.11	1.19	1.16
Melanoma	0.46	0.47	0.46	0.80	1.04	0.77
Myeloma	-	0.32	0.30	-	1.29	1.13
Neuro	0.36	0.41	0.37	0.69	0.98	0.75
Thoracic	0.34	0.35	0.34	1.20	0.71	1.19
Thyroid	-	0.41	0.39	-	1.42	1.09

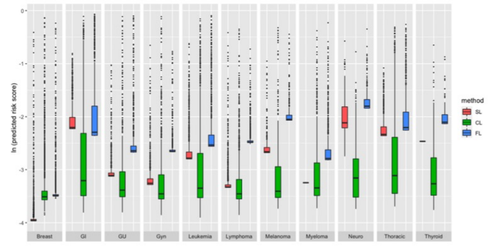

Figure 1. Distribution of log-transformed predicted risk estimates of mortality across cancer types.

Conclusion

The proposed method groups the parameters based on their similarities in the coefficients, which is a good alternative approach to the multiple hypothesis testing in the presence of numerous subgroups. Such fusion is also expected to achieve parsimony of model parameters. Our method has also shown an improved empirical performance, especially when it comes to calibration.

References

[1] D. P. Bertsekas. Nonlinear programming. *Athena Scientific, Belmont, MA*, 1999.

[2] R. B. Parikh, C. Manz, C. Chivers, and et al. Machine learning approaches to predict 6-month mortality among patients with cancer. *JAMA Network Open*, 2(10):e1915997, 2019.

[3] Lu Tang and Peter X.K. Song. Fused lasso approach in regression coefficients clustering – learning parameter heterogeneity in data integration. *Journal of Machine Learning Research*, 17(113):1–23, 2016.