The Annals of Applied Statistics
2019, Vol. 13, No. 4, 2637–2661
https://doi.org/10.1214/19-AOAS1292
© Institute of Mathematical Statistics, 2019

HIERARCHICAL INFINITE FACTOR MODELS FOR IMPROVING THE PREDICTION OF SURGICAL COMPLICATIONS FOR GERIATRIC PATIENTS

BY ELIZABETH LORENZI[*], RICARDO HENAO[†] AND KATHERINE HELLER[†]

Berry Consultants[*] *and Duke University*[†]

Nearly a third of all surgeries performed in the United States occur for patients over the age of 65; these older adults experience a higher rate of postoperative morbidity and mortality. To improve the care for these patients, we aim to identify and characterize high risk geriatric patients to send to a specialized perioperative clinic while leveraging the overall surgical population to improve learning. To this end, we develop a hierarchical infinite latent factor model (HIFM) to appropriately account for the covariance structure across subpopulations in data. We propose a novel Hierarchical Dirichlet Process shrinkage prior on the loadings matrix that flexibly captures the underlying structure of our data while sharing information across subpopulations to improve inference and prediction. The stick-breaking construction of the prior assumes an infinite number of factors and allows for each subpopulation to utilize different subsets of the factor space and select the number of factors needed to best explain the variation. We develop the model into a latent factor regression method that excels at prediction and inference of regression coefficients. Simulations validate this strong performance compared to baseline methods. We apply this work to the problem of predicting surgical complications using electronic health record data for geriatric patients and all surgical patients at Duke University Health System (DUHS). The motivating application demonstrates the improved predictive performance when using HIFM in both area under the ROC curve and area under the PR Curve while providing interpretable coefficients that may lead to actionable interventions.

1. Introduction. Surgical complications arise in 15% of all surgeries performed and increases up to 50% in high-risk surgeries (Healey et al. (2002)). Surgical complications are associated with decreased quality of life to patients and also incur significant costs to the health system. Efforts to address this problem are increasing nationwide with a focus on enhancing preoperative and perioperative care for high-risk and high-cost patients (Desebbe et al. (2016)). Duke University Health System (DUHS) began the Perioperative Optimization of Senior Health (POSH) program, an innovative care redesign that uses expertise from geriatrics, general surgery and anesthesia to focus on the aspects of care that are most influential for the geriatric surgical population.

Received July 2018; revised May 2019.

Key words and phrases. Bayesian factor model, nonparametrics, transfer learning, hierarchical modeling, surgical outcomes, health care.

Nearly a third of all surgeries performed in the United States occur for people over the age of 65. Furthermore, these older adults experience a higher rate of postoperative morbidity and mortality (Etzioni et al. (2003), Hanover (2001)). Complications for older adults may also lead to slower recovery, longer postoperative hospital stays, more complex care needs at discharge, loss of independence and high readmission rates (Speziale et al. (2011), Raval and Eskandari (2012)). The established predictors of poor outcomes, such as age, presence of comorbidities and the type of surgical procedure performed, are important predictors for all patient populations, including the geriatric population. However, other factors such as functional status, cognition, nutrition, mobility and recent falls are less routinely collected factors that are highly correlated with surgical risk among older adults (Jones et al. (2013)). Based on this research, POSH developed a heuristic to determine which patients to refer from the surgery clinic visit to their specialized clinic. The heuristic is defined as all patients 85 or older or patients that are 65 or older with cognitive impairment, recent weight loss, multimorbid or polypharmacy. However, the heuristic identifies about a quarter of the volume of all invasive surgical encounters, which results in more patient visits than POSH can accommodate.

Our goal is to identify and characterize high-risk geriatric patients who are undergoing an elective, invasive surgical procedure to send to the specialized POSH clinics. We leverage the larger surgical population at DUHS to improve learning, using data derived from electronic health records (EHR). We develop a sparse multivariate latent factor model to learn an underlying latent representation of the data that adjusts for the differences between geriatric surgical patients and all other surgical patients. Our approach builds on the framework introduced by West (2003) and extended by Avalos-Pacheco, Rossell and Savage (2018), Bhattacharya and Dunson (2011), Carvalho et al. (2008a), Carvalho et al. (2006), Ročková and George (2016) all of which consider nonparametric priors to incorporate flexibility in learning the number of factors, either for unsupervised or supervised learning. In addition, Chen et al. (2010), Murphy, Gormley and Viroli (2017) have proposed sparse factor models in mixture model contexts.

Working with high-dimensional EHR data introduces the problems of noisiness, sparsity and multicollinearity among the covariates. We therefore model the factor loadings matrix as sparse, assuming that only a few variables are related to the factors and thus the factors represent a parsimonious representation of the data. This modeling approach serves as an exploratory view of underlying phenotypes of the geriatric population compared to the full population that can guide surgeons in deciding which profile of patients would most benefit from interventions. This is similar to recent work by Ni, Mueller and Ji (2018), where they propose a categorical matrix factorization method to infer latent diseases from EHR data, though only in an unsupervised manner, as a phenotype discovery tool. In our work we additionally incorporate response variables, using our learned phenotypes to predict post-operative surgical complications with high accuracy through a supervised approach.

We focus on modeling the covariance structure of different subpopulations to adjust for the idiosyncratic variations and covariations of each subpopulation. Latent factor models aim to explain the dependence structure of observed data through a sparse decomposition of their covariance matrix. Specifically, factor models decompose the covariance of the observed data of p dimensions, Ω, as $\Lambda\Lambda^T + \Sigma$, where Λ is a $p \times k$ loadings matrix that defines the relationships between each covariate and k latent factors, and Σ is a $p \times p$ diagonal matrix of idiosyncratic variances. These models are often used in applications in which the latent factors naturally represent some hidden features such as psychological traits or political ideologies. Others find utility in their use as a dimensionality reduction tool for prediction problems with large p and small n (West (2003)). However, our data, derived from noisy EHR, call for flexibility beyond the common factor model to better handle the complex structure of the subpopulations we consider and to induce strong sparsity that can improve predicting outcomes with very low prevalence. A key contribution of this work is the development of the sparse factor model into a transfer learning approach, where we utilize data from a larger source population to improve learning in a target population, in our case the geriatric patients qualified for POSH through their established heuristic. Similar work by Seo, Goldschmidt-Clermont and West (2007) uses sparse factor models to tie together gene expression data from two populations: mice and men, tackling a similar problem of sharing information between two populations though through a different modeling framework.

Our proposed transfer learning approach places hierarchical priors on the factor loadings matrix. In this setting we define two groups or subpopulations: the POSH heuristic defined cohort of patients and the remaining invasive procedures occurring at Duke from the entire patient population over the age of 18. The motivating reason for selecting the subpopulation, as determined through the heuristic, is to align with current interventions currently deployed in the POSH clinic. The interventions at POSH, such as management of comorbidities, reduction of polypharmacy, enhancement of mobility and nutrition and delirium risk mitigation, are targeted to geriatric patients and their previously studied risk factors (McDonald et al. (2018)). Therefore, to align with the goal of the project, we use this known targeted subpopulation to assure that the interventions are effective for the patients triaged to the clinic.

There are inherent differences between these two populations. In Figure 1 we present the t-distributed Stochastic Neighbor Embedding (t-SNE) representation of the EHR surgical database of all invasive procedures at Duke University Health System (DUHS) between January 2014–January 2017, with samples of 10,000 from the geriatric population that meets the POSH heuristic requirements and 10,000 from all other surgical encounters. The figure shows patient substructure in the data with a clear difference in the two populations. While there is some overlap between the two populations, it is clear geriatric patients have a different covariate space compared to the overall population, and therefore should be modeled appropriately.

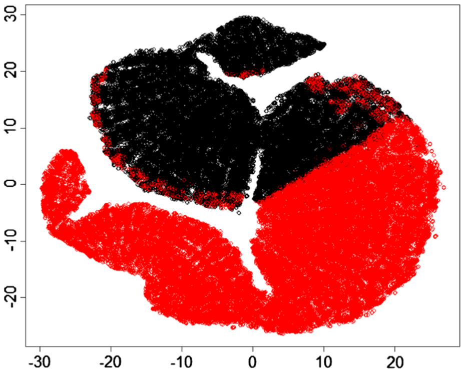

FIG. 1. *t-SNE representation of EHR data from Duke University that meets the POSH heuristic* (*red*) *and full patient populations* (*black*), *using samples of size* 10,000 *for each group. Displays low-dimensional projection of full data.*

Modeling these disparate populations requires proper adjustments. Therefore, we introduce a hierarchical infinite prior on the factor loadings matrix which learns the proper number of factors needed in each group's factor model while still sharing information across groups to aid in learning for the smaller subpopulation. The hierarchical infinite prior for the factor loadings matrix is derived from the Hierarchical Dirichlet Process (HDP), a nonparametric model most commonly used within a mixture model, where one may be interested in learning clusters among multiple groups (Teh et al. (2006)). The hierarchical infinite prior combines ideas from sparse Bayesian factor models with the hierarchical grouping characteristics of an HDP mixture model, aiming to share information between subpopulations while capturing the underlying cluster structure, similar to the HDP. We also aim to decompose the sparse covariance structure of our data to model directly the main source of variation between groups, as in a *hierarchical* factor model. We therefore place a hierarchical prior on the loadings matrix of our factor model, Λ, that flexibly captures the underlying structure of our data across populations.

Section 2 provides necessary background, a detailed overview of our proposed hierarchical infinite factor model (HIFM) utilizing the HDP and resulting properties of the prior showing that it is well defined and result in a semidefinite covariance matrix. Section 3 presents results in simulations, portraying the properties in model selection, prediction and interpretability. Section 4 discusses the derived EHR data used to predict surgical complications and reviews the results. Section 5 concludes and presents future directions for continued effort.

2. Hierarchical infinite factor model.

2.1. *Primitives.* The standard Bayesian latent factor model relates observed data, $\boldsymbol{x}_i$, to an underlying k-vector of random variables, $\boldsymbol{f}_i$, using a standard k-

factor model for each observation, $i \in 1, \dots, n$ (Lopes and West (2004)):

$$x_i \sim N(\Lambda f_i, \Sigma), \tag{2.1}$$

where $\boldsymbol{x}_i$ is a p-dimensional vector of covariates, assumed continuous, Λ is the $p \times k$ factor loadings matrix where the jth row is distributed $\boldsymbol{\lambda}_j \sim N(0, \phi^{-1} I_k)$. The k-dimensional factors, $\boldsymbol{f}_i$, are independent and identically distributed as $\boldsymbol{f}_i \sim N(0, I_k)$, and $\Sigma = \operatorname{diag}(\sigma_1^2, \dots, \sigma_p^2)$ is a diagonal matrix that reduces to a set of p independent inverse gamma (IG) distributions, with $\sigma_j^2 \sim \mathrm{IG}(a, b)$ for $j = 1, \dots, p$. Conditioned on the factors, the observed variables are uncorrelated. Dependence between these observed variables is induced by marginalizing over the distribution of the factors, resulting in the marginal distribution, $\boldsymbol{x}_i \sim N_p(0, \Omega)$ where $\Omega = V(\boldsymbol{x}_i | \Lambda, \Sigma) = \Lambda\Lambda^T + \Sigma$. Note that there is not an identifiable solution to the above specification and therefore the decomposition of Ω is not unique. However, for problems involving covariance estimation and prediction, this requirement is not needed, and therefore we do not impose any constraints on the model. This allows us to construct a more flexible parameter-expanded loadings matrix.

We propose a hierarchical stick-breaking prior, motivated by the HDP. The HDP is hierarchical, nonparametric model in which each subpopulation is modeled with a DP, where the base measure are DP themselves. The DP, $DP(\alpha_0, G_0)$, is a measure on (probability) measures, where $\alpha_0 > 0$ is the concentration parameter, and G_0 is the base probability measure (Ferguson (1973)). A draw from a DP is formulated as $G = \sum_{h=1}^{\infty} \pi_h \delta_{\phi_h}$, where ϕ_h are independent random variables from G_0 and δ_{ϕ_h} are atom locations at ϕ_h, and π_h are the "stick-breaking" weights that depend on the parameter α_0 (Sethuraman (1994)). The HDP is a hierarchical model in which the base measure for the children DP are DP themselves, such that

$$\begin{aligned} G_0 | \alpha_0, H &\sim DP(\alpha_0, H) \\ G_l | \alpha_l, G_0 &\sim DP(\alpha_l, G_0), \quad \text{for each } l. \end{aligned}$$

This results in each group sharing the components or atom locations, $\boldsymbol{\phi}$, while allowing the size of the components to vary per group.

2.2. *Proposed model.* Now, consider a $p \times \infty$ loadings matrix, Λ_0, weighted by the stick-breaking weights of an HDP, such that each population has a unique loadings matrix defined by population specific weights, $\boldsymbol{\pi}_l$, where $\Lambda_l = [\sqrt{\pi_{l1}}\lambda_{01}, \sqrt{\pi_{l2}}\lambda_{02}, \dots]$. The population specific loadings matrix becomes a weighted version of a shared global loadings matrix. The Bayesian factor model prior specification assumes independent rows and columns, so an element in row j and column h from Λ_0, λ_{0jh} is distributed as a zero-mean normal distribution. Multiplying λ_{0jh} by $\sqrt{\pi_{lh}}$ results in $\sqrt{\pi_{lh}}\lambda_{0jh} \sim \mathrm{N}(0, \pi_{lh}\phi^{-1})$. This now mimics the formulation of a scale mixture with the full specification shown in (2.2) where we represent the prior in the finite case for clarity of the scale mixture specification. We let $D_{lj} = \operatorname{diag}(\pi_{l1}/\phi_{j1}, \dots \pi_{lk}/\phi_{jk})$:

$$\begin{aligned} \boldsymbol{\lambda}_{lj}|\boldsymbol{\pi}_l &\sim N(0, D_{lj}), \\ \boldsymbol{\pi}_l|\boldsymbol{\pi}_0 &\sim \operatorname{Dir}(\alpha_l \boldsymbol{\pi}_0), \\ \boldsymbol{\pi}_0 &\sim \operatorname{Dir}(\alpha_0/k, \dots, \alpha_0/k), \\ \phi_{jh} &\sim \operatorname{Gamma}(\tau/2, \tau/2). \end{aligned} \tag{2.2}$$

The nonparametric process representation is recovered if we let $k \to \infty$ (Teh et al. (2006)). We continue with the finite truncation of the model, which is known to be "virtually indistinguishable" from the full process (Ishwaran and James (2001), Ishwaran and James (2002)), with k^* as a large *upper bound* for the number of factors. For convenience, we continue to use the notation k where k is sufficiently large. While we focus on the finite version of the model for computational reasons, the underlying infinite process provides the adaptability of learning the number of factors, a key feature of the proposed model. Note that it is common in the literature to suggest a truncated version of nonparametric factor models (Bhattacharya and Dunson (2011), Ročková and George (2016), Ni, Mueller and Ji (2018)).

We set the scale parameter in the loadings matrix, ϕ_{jh}, constant across populations and distributed gamma in such a way that marginally ϕ_{jh} results in a t-distribution with τ degrees of freedom, resulting in a heavy tailed distribution.

The Dirichlet distribution can be decomposed into a set of k independent gamma distributions, such that $w_h \sim \operatorname{Gamma}(\alpha_h, 1)$ for $h = 1, \dots, k$ and $S := (w_1 + \cdots + w_k)$, then $(w_1/S, \dots, w_k/S) \sim \operatorname{Dir}(\alpha_1, \dots, \alpha_k)$. We show this for the finite case, but the same is true in the infinite limit where the Gamma distribution becomes a Gamma process. To induce a closed-form posterior for our proposed prior, we use k unnormalized Gamma draws, $\boldsymbol{w}_l$, instead of a draw from a Dirichlet, $\boldsymbol{\pi}_l$. The resulting hierarchical prior is specified below in (2.3) where, now, we let $D_{lj} = \operatorname{diag}(w_{l1}/\phi_{j1}, \dots, w_{lk}/\phi_{jk})$:

$$\begin{aligned} \boldsymbol{\lambda}_{lj}|\boldsymbol{w}_l, \boldsymbol{\phi}_j &\sim N(0, D_{lj}), \\ w_{lh}|\pi_{0h} &\sim \operatorname{Gamma}(\alpha_l \pi_{0h}, 1), \quad \forall h \in 1, \dots, k, \\ \boldsymbol{\pi}_0 &\sim \operatorname{Dir}(\alpha_0/k, \dots, \alpha_0/k), \\ \phi_{jh} &\sim \operatorname{Gamma}(\tau/2, \tau/2), \\ \sigma_j^2 &\sim \operatorname{IG}(a, b). \end{aligned} \tag{2.3}$$

Our prior formulation does not require that $\boldsymbol{w}_l$ sums to one, as is the case in a Dirichlet draw. We want the "rich gets richer" behavior of the HDP that results in many of the stick-breaking weights being approximately zero, signifying the absence of those clusters. By unnormalizing the weights the same scaling occurs where some weights will be much smaller than others, but now the magnitude is not bounded. This acts as a model-shrinkage tool for shrinking factors not needed to describe the distribution of group l. We prove a subsequent result in Section 2.4, showing that a loadings matrix with infinite columns results in a finite loadings matrix and covariance structure. The most prominent difference between the HDP and our weighting scheme is that we are not drawing from a discrete measure, instead we use the properties inherent in the stick-breaking process of the sampling proportions to weigh the importance of factors in our model.

As discussed in Polson and Scott (2011), scale mixtures should meet two criteria: first, a local scale parameter should have heavy tails to detect the signal, and, second, a global scale parameter should have substantial mass at zero to handle the noise. Marginalizing over the weights, $\boldsymbol{w}_l$, the resulting distribution of Λ relates to the normal gamma shrinkage prior discussed in Caron and Doucet (2008). To avoid over shrinking the nonzero loadings, we also define a $p \times k$ matrix of local scale parameters ϕ drawn elementwise from a gamma distribution that is constant across populations. This adds an additional source of sharing of information or transfer learning. For example, if an element of the loadings matrix is near zero with small variance, then the signal will also be similar for other subpopulations.

2.3. *Hierarchical latent factor regression.* We utilize the hierarchical infinite factor model to relate the observed covariates to response variables. For each $\boldsymbol{x}_i$, we have a corresponding response or a p_y-dimensional vector of responses, $\boldsymbol{y}_i \in \{0, 1\}$. Let $Z = \{Y, X\}$ represent the full data, and the model in 2.1 simply replaces the x_i with z_i. We concatenate $[\boldsymbol{f}_i, 1]$ and learn an additional column of the loadings matrix. The $k + 1$ column of the loadings matrix now serves as an intercept in the model for each covariate.

The posterior predictive distribution is easily obtained by solving

$$\begin{aligned} &f(y_{n+1}|z_1, \dots, z_n, x_{n+1}) \\ &\quad = \int f(y_{n+1}|x_{n+1}, \Theta)\pi(\Theta|z_1, \dots, z_n)\, d\Theta. \end{aligned}$$

The joint model implies that $E(y_i|x_i) = x_i'\beta_l$ with covariance matrix $\Omega_{l,YX}$, where $\Omega_{l,YX}$ is a partitioned covariance matrix defined by the rows and columns corresponding to Y and X. The resulting coefficients, $\beta_l = \Omega_{l,XX}^{-1}\Omega_{l,YX}$, are found by correctly partitioning the covariance matrix, Ω_l. This then results in the true group-specific regression coefficients of Y on X.

In our application the data are both binary and continuous, with all outcomes being binary indicators of surgical complications. Therefore, we extend this method to deal with this data structure by using the common probit transformation (Albert and Chib (1993)). We choose this transformation due to its ease in computation and implementation. It is also commonly seen throughout the literature as a promising way of dealing with mixed data, with others using this type of transformations in latent variable models (McParland et al. (2014, 2017)). With the probit transformation we convert our binary data to the real line, where it now mimics a Gaussian likelihood, as the continuous variables do under our model specifications, except in this case we do not learn the idiosyncratic noises, Σ and instead set those to 1.

The resulting factor scores represent a transformed feature space of our data that aim to minimize the distributional differences between the populations. Therefore, we proceed with prediction by learning factor scores for the held-out test set of interest. Specifically, we will draw $p(f_i|x_i, \Lambda_{XX}, \Sigma_{XX})$ for each i in the testing set from the defined full conditional for f_i where we subset the learned parameters appropriately to match the testing predictors.

2.4. *Properties of the shrinkage prior.* We let $\Pi_\Lambda \otimes \Pi_\Sigma$ be the prior specification defined in (2.3). Because Π_Λ defines the prior on the infinite dimensional loadings matrix, we must assure that a draw from the prior is well defined and that the elements of the $\Lambda\Lambda^T$ are finite for a semidefinite covariance matrix. As shown in Bhattacharya and Dunson (2011), we can define a loadings matrix, Λ, with infinitely many columns while keeping $\Lambda\Lambda^T$'s entries finite. We follow the steps taken in their paper to prove similar properties for our hierarchical infinite factor loadings prior.

We first define Θ_Λ as the collection of matrices Λ with p rows and infinite number of columns, such that the $p \times p$ matrix, $\Lambda\Lambda^T$, results in all finite entries:

$$\begin{aligned}\Theta_\Lambda = \Bigl\{ &\Lambda = (\lambda_{jh}), j = 1, \dots, p, \\ &h = 1, \dots, \infty, \max_{1 \le j \le p} \sum_{h=1}^{\infty} \lambda_{jh}^2 < \infty \Bigr\}.\end{aligned} \tag{2.4}$$

The entries of $\Lambda\Lambda^T$ are finite if and only if the condition in (2.4) is satisfied, which is possible using the Cauchy–Schwarz inequality and proved in the Supplement A (Lorenzi, Henao and Heller (2019)). All proofs for subsequent properties are shown in Supplement A (Lorenzi, Henao and Heller (2019)).

Next, let Θ_Σ denote the $p \times p$ diagonal matrices with nonnegative entries, let Θ denote all $p \times p$ positive semidefinite matrices and allow $g : \Theta_\Lambda \times \Theta_\Sigma \to \Theta$ corresponding to $g(\Lambda, \Sigma) = \Lambda\Lambda^T + \Sigma$. We next define Proposition 1 to show that our prior is an element of $\Theta_\Lambda \times \Theta_\Sigma$ almost surely. This reduces to a proof of $\Pi_\Lambda(\Theta_\Lambda) = 1$ under the independence assumption on $\Theta_\Lambda \times \Theta_\Sigma$ where Θ_Σ is well defined as a product of p inverse-gamma distributions.

PROPOSITION 1. *If* $(\Lambda, \Sigma) \sim \Pi_\Lambda \otimes \Pi_\Sigma$, *then* $\Pi_\Lambda \otimes \Pi_\Sigma(\Theta_\Lambda \times \Theta_\Sigma) = 1$.

We also show that the resulting posterior distribution of the marginal covariance, $\Omega = \Lambda\Lambda^T + \Sigma$, is weakly consistent by proving Theorem 1, defined below:

THEOREM 1. *Fix $\Omega_0 \in \Theta$. For any $\epsilon > 0$, there exists $\epsilon^* > 0$ such that*

$$\{\Omega : d_\infty(\Omega, \Omega_0) < \epsilon^*\} \subset \{\Omega : K(\Omega_0, \Omega) < \epsilon\}.$$

Our infinite hierarchical prior meets these properties for each group's estimated covariance by first showing that the prior has large support and, therefore, places positive probability in ϵ-neighborhoods around any covariance matrix.

Lastly, we make an argument that the resulting covariance decomposition mimics the results from an HDP mixture model with cluster-specific covariances. For group l, Ω_l is the population-specific covariance structure of the data, X, where $\Omega_l = \Lambda_l \Lambda_l^T + \Sigma_l$. If we rewrite Λ_l as $(\Lambda_0 W_l^{1/2})$ where W_l is diagonal matrix of elements $\boldsymbol{w}_l$, we see that the resulting decomposition is $(\Lambda_0 W_l \Lambda_0^T) + \Sigma$. We then can reformulate this as a sum up to k (with $k \to \infty$), resulting in a linear combination of rank-1 covariance matrices:

$$\Omega_l = \Lambda_l \Lambda_l^T + \Sigma_l \tag{2.5}$$

$$= (\Lambda_0 W_l \Lambda_0^T) + \Sigma_l \tag{2.6}$$

$$= \sum_{h=1}^{k} w_h (\boldsymbol{\lambda}_{0h} \boldsymbol{\lambda}_{0h}^T) + \Sigma_l. \tag{2.7}$$

2.5. *Inference.* We propose a Markov chain Monte Carlo (MCMC) scheme with almost all closed-form updates, and provide some suggested updates to allow for faster computation. We truncate the loadings matrix to have $k^* < p$.

We derive a Gibbs sampler where we draw from the full conditional posteriors. Most posterior updates are derived from conjugate relationships; however, the parameters for the unnormalized HDP are not conjugate. The weight parameters $\boldsymbol{w}_l$ are updated with a closed form draw from the generalized inverse-gaussian (GIG) distribution for each hth element of w_l:

$$w_{lh} | \lambda_l, \pi_0, \alpha_l \sim \mathrm{GIG}(p = p_{w_{lh}}, a = a_{w_{lh}}, b = b_{w_{lh}})),$$

where $p_{w_{lh}} = \alpha_l \pi_h^0 - p/2$, $a_{w_{lh}} = 2$, and $b_{w_{lh}} = (\boldsymbol{\lambda}_{lh}' \Phi_h \boldsymbol{\lambda}_{lh})$. $\Phi_h = \mathrm{diag}(\phi_{h1}, \ldots, \phi_{hp})$.

To update $\boldsymbol{\pi}_0$, we use a Metropolis–Hastings step within the Gibbs sampler using a gamma proposal with normalization to mimic the Dirichlet distribution. This is done in two steps: First, we propose $\theta_{0h}^* \sim \mathrm{Gamma}(\theta_{0h}^{t-1} \cdot C, C)$, which gives a mean of θ_{0h}^{t-1} and a variance of θ_{0h}^{t-1}/C, which allows tuning using the constant, C. We then normalize the $\boldsymbol{\theta}_0^*$, such that $\boldsymbol{\pi}_0^* = \frac{\boldsymbol{\theta}_0^*}{\sum_{h=1}^k \boldsymbol{\theta}_{0h}^*}$, and accept π_0^* based on the acceptance ratio

$$A(\boldsymbol{\pi}_0^* | \boldsymbol{\pi}_0^{t-1}) = \min\left(1, \frac{P(\boldsymbol{\pi}_0^* | w_1, \ldots, w_l)}{P(\boldsymbol{\pi}_0^{t-1} | w_1, \ldots, w_l)} \frac{g(\boldsymbol{\pi}_0^{t-1} | \boldsymbol{\pi}_0^*)}{g(\boldsymbol{\pi}_0^* | \boldsymbol{\pi}_0^{t-1})}\right).$$

The acceptance ratio for $\boldsymbol{\pi}_0$ helps to serve as a hyperparameter check for your selected α_0 and α_l. With poor choices of k and α_0 and α_k, the proposals for $\boldsymbol{\pi}_0$ will be poor and few will be accepted. We recommend acceptance ratios between 0.2 and 0.5, as is common in the Bayesian literature.

All remaining updates from the Gibbs sampler are presented in Supplement B (Lorenzi, Henao and Heller (2019)). To speed the computation time of this sampler, we parallelize the updates for the factors $\boldsymbol{f}_i$ and the probit transformations of $\boldsymbol{x}_i$. Because we assume each row of $\boldsymbol{f}_i$ and $\boldsymbol{x}_i$ are independent and identically distributed (within each population), we are able to split this update using parallel methods and speed up each iteration by a factor of the number of cores or computing resources present.

3. Simulations.

We next evaluate our approach through synthetic data and compare to baseline methods, Lasso and elastic net regressions (Tibshirani (1996), Zou and Hastie (2005)). Lasso is a commonly used penalized regression model used for variable selection that excels when working with sparse, correlated data while providing interpretable coefficients that provide insight into the underlying relationships between covariates and outcomes. Elastic net pairs Lasso with ridge regression to share the benefit of both variable selection and regularization and often results in grouping effects among correlated coefficients. When considering comparison methods, we selected models that would commonly be deployed for the problem of predicting surgical complications with an interpretable model. In addition, we looked to methods that had available code. The Lasso and elastic net models were chosen due to their ability to accurately predict binary outcomes while providing interpretable coefficients for both binary and continuous variables. The goal of these analyses is to demonstrate HIFM's capabilities as an interpretable and flexible factor model that excels at prediction. To this end, we design two simulation studies. The first considers data generated from a slight deviation from the underpinning model, with variations of the dimension considered, ranging from $p = 50$ to $p = 250$. The second setting follows that of Bhattacharya and Dunson (2011) to generate a sparse covariance matrix with extensions to the multisubpopulation setting where there are some relationship between the subpopulations.

First setting: We simulate data, z_i, for $i = 1, \dots, 1000$ from a p-dimensional normal distribution, with zero mean and covariance equal to $\Omega_l = \Lambda_l \Lambda_l^T + \Sigma_l$. We simulate with two populations where 400 observations are within $l = 1$, our target. We draw the jth row of λ_{lj} from a $\mathrm{N}(0, D_{lj}^{-1})$ where $D_{lj}^{-1} = \mathrm{diag}(\boldsymbol{\phi}_j / \boldsymbol{w}_l)$ is a $k \times k$ diagonal matrix. We draw each ϕ_{jh} for $j \in 1, \dots, p$ and $h \in 1, \dots, k$ from a $\mathrm{Gamma}(\tau/2, \tau/2)$ where $\tau = 3$, $w_{lh} \sim \mathrm{Gamma}(\alpha_l \pi_0, 1)$, and π_0 from a $\mathrm{Dir}(\boldsymbol{\alpha}_0 / k)$ with hyperparameters set to $\alpha_l = \alpha_0 = 15$ to induce approximately uniform clusters. We set the first row of the loadings matrix that corresponds to the outcome to zero and randomly select two locations and fill in with a 1 and -1 to induce further sparsity. This adjustment in the simulation aims to induce even stronger sparsity between the response variable and the predictors, as well as to create a generative process of the data that is not exactly that of our model. We draw the diagonal of Σ from IG(1, 0.33) with prior mean equal to 3.

Second setting: For the second study, we simulate data, z_i, for $i = 1, \dots, 1000$ from a p-dimensional normal distribution with zero mean and covariance equal to $\Omega_l = \Lambda_l \Lambda_l^T + \Sigma_l$. We simulate with two populations, where 400 observations are within $l = 1$, our target. In generating the Λ_l and Σ_l we mimic the simulation scheme presented in Bhattacharya and Dunson (2011) with slight moderations. We generate a global loadings matrix, Λ_0, that will be the mean of each subpopulation's Λ_l. For each column of Λ_0, the number of nonzero elements is chosen linearly between $2k$ and $k + 1$, with the zeros randomly allocated in each column. The nonzero elements in Λ_0 are drawn independently from N(0, 5). Then, for each subpopulation's loading matrix, Λ_l, the nonzero elements are drawn from a normal distribution centered at the nonzero elements in Λ_0 with additional standard deviation of 1. For each Λ_l, we set the first row of the loadings matrix that corresponds to the outcome, y, to zero and randomly select two locations and fill in with a 1 and -1 to induce further sparsity. We draw the diagonal of each Σ_l^{-1} from Gamma(1, 0.33) with prior mean equal to 3 for each population independently.

For both settings, we compare three different choices of p, 50, 100 and 250 with the true number of factors $k = 10$. We use the default choice of $5\log(p)$ as the starting number of factors for each simulation run to select an upper bound on k. For each run, we sampled from the Gibbs sampler for 4000 iterations and remove 2000 iterations for burnin and thinned every fifth iteration. We show two examples: the first with all continuous data as described above, and the second converts the Gaussian simulated data into binary columns using the probit transformation and a random binomial. We convert the first $p/2$ columns, including the outcome, into binary variables for both simulations cases with varying p.

We repeat the simulations 50 times and evaluate: (1) the prediction performance using an out of sample test set, (2) the precision of the estimated coefficients, and (3) the estimation of the number of factors. We calculate the prediction accuracy for continuous outcomes with mean squared error (MSE) and the binary outcome using area under a receiver operator characteristic curve (AUC), by reporting the median, minimum and max from the 50 runs. We compare the HIFM model to elastic net and Lasso trained with two different covariate specifications. The first Lasso uses all covariates as main effects and ignores the subpopulation, and the second incorporates a random slope per subpopulation through interactions, which we call a hierarchical Lasso. To tune these models, we use 10-fold cross validation. For Lasso we use the cv.glmnet function from the package glmnet with their default tuning settings. For elastic net we cross validate with a grid of 30 parameter settings, where alpha ranges between 0 to 1 in increments of 0.1, and lambda ranges between 0.001 and 1e$-$5 (using the default tuning grid for lambda from cv.glmnet).

Tables 1 and 2 display the results from the two simulation settings, respectively. For both simulation settings, when all data are continuous, the hierarchical infinite factor model achieves superior predictive performance compared to elastic

TABLE 1

Predictive performance in simulation study for first simulation setting cases. Average, minimum and maximum performance is presented across 50 *simulations. Mean squared error* (*MSE*) *is calculated for continuous outcome simulations* (*where smaller is better*). *Area under receiver operator characteristic curve* (*AUC*) *is reported for binary outcomes* (*where closer to* 1 *is better*). (*EN-elastic net, L-Lasso, HL-hierarchical Lasso*)

	MSE				AUC			
	HIFM	EN	L	HL	HIFM	EN	L	HL
(50, 10)								
Mean	**0.71**	0.74	0.88	0.91	**0.83**	0.81	0.80	0.81
Min	0.08	0.11	0.12	0.12	0.59	0.62	0.63	0.61
Max	4.81	4.58	4.91	5.07	0.93	0.93	0.91	0.92
(100, 10)								
Mean	**0.88**	0.89	1.02	1.06	**0.83**	0.81	0.78	0.79
Min	0.08	0.09	0.11	0.12	0.52	0.54	0.55	0.52
Max	5.54	5.47	5.44	5.49	0.95	0.94	0.89	0.91
(250, 10)								
Mean	**1.70**	**1.70**	1.81	1.82	0.84	0.85	**0.85**	0.85
Min	0.12	0.15	0.19	0.17	0.49	0.55	0.53	0.47
Max	20.24	19.85	20.55	20.74	0.93	0.94	0.945	0.944

net, Lasso and a hierarchical Lasso. Tables 1 and 2 also display the AUC calculated across 50 simulations with binary outcomes where again HIFM outperforms the alternative models. The baselines provide two gold standards in sparse regression modeling. Elastic net performs slightly better in prediction tasks compared to Lasso and hierarchical Lasso, and hierarchical Lasso does improve over Lasso, suggesting the interactions help to better capture the group effects. HIFM improves predictive performance for both continuous and binary outcomes, compared to the alternatives. However, in $p = 250$ under the first simulation scheme, our predictive performance is worse than Lasso in terms of AUC, albeit by a very slim margin (0.84 HIFM versus 0.85 Lasso). We see the performance does equally well between the two data simulation schemes, suggesting in sparse settings with related populations the model is able to perform with high prediction accuracy.

Tables 3 and 4 display resulting accuracy of the learned coefficients for each population across method. Coefficients from HIFM are derived from transforming the partitioned covariance matrix of the learned model. We compare the results of HIFM learned regression coefficients to those learned by Lasso with and without interactions and elastic net. We display the results for the simulation of $p = 100$ and $k = 10$ for both continuous and binary outcomes and averaged over 50 iterations. Similar patterns occurred in the smaller covariate simulation cases, where $p = 50$ and the larger covariate case when $p = 250$; therefore, we do not report these additional results. The hierarchical Lasso improves the model fit compared

TABLE 2

Predictive Performance in simulation study for second simulation setting cases. Average, minimum, and maximum performance is presented across 50 *simulations. Mean squared error* (*MSE*) *is calculated for continuous outcome simulations* (*where smaller is better*). *Area under receiver operator characteristic curve* (*AUC*) *is reported for binary outcomes* (*where closer to* 1 *is better*). (*EN-elastic net*, *L-Lasso*, *HL-hierarchical Lasso*)

	MSE				AUC			
	HIFM	EN	L	HL	HIFM	EN	L	HL
(50, 10)								
Mean	**0.76**	0.78	0.89	0.89	**0.82**	0.80	0.78	0.80
Min	0.07	0.09	0.12	0.12	0.61	0.59	0.59	0.59
Max	3.97	3.96	4.28	4.11	0.94	0.92	0.89	0.91
(100, 10)								
Mean	**1.98**	2.02	2.10	2.10	**0.83**	0.81	0.78	0.79
Min	0.07	0.09	0.20	0.18	0.52	0.54	0.55	0.52
Max	38.65	37.60	36.99	37.26	0.95	0.94	0.89	0.91
(250, 10)								
Mean	**3.00**	3.16	3.19	3.18	**0.82**	0.81	0.81	0.81
Min	0.09	0.10	0.11	0.10	0.52	0.53	0.54	0.55
Max	70.00	71.97	71.12	71.12	0.91	0.91	0.94	0.92

to regular Lasso, providing evidence that modeling these data hierarchically aids in coefficient estimation. Compared to Lasso and elastic net, HIFM captures the true coefficients with greater accuracy for both populations. Interestingly, elastic net performs much worse in the estimation of regression coefficients, compared to

TABLE 3

Performance in estimating regression coefficients in first simulation study. We report results with $p = 100$, $k = 10$ *for* 50 *simulations for both continuous and binary examples*, *showing mean squared error* ($\times 10^3$) *of estimated coefficients compared to true simulated coefficients*

	Continuous outcomes				Binary outcomes			
	HIFM	EN	L	HL	HIFM	EN	L	HL
Pop 1:								
Median	**0.02**	0.11	0.13	0.09	**0.05**	4.96	0.73	0.27
Min	0.00	0.06	0.06	0.04	0.01	0.10	0.07	0.05
Max	1.82	1.17	0.25	0.19	10.94	76.40	8.74	7.04
Pop 2:								
Median	**0.89**	0.93	0.95	0.97	**1.11**	11.90	2.55	1.93
Min	0.04	0.07	0.05	0.03	0.05	0.08	0.05	0.05
Max	5.76	0.29	0.20	0.24	19.26	76.07	8.72	6.08

TABLE 4

Performance in estimating regression coefficients in second simulation study. We report results with $p = 100$, $k = 10$ *for* 50 *simulations for both continuous and binary examples, showing mean squared error* $(\times 10^3)$ *of estimated coefficients compared to true simulated coefficients*

	Continuous Outcomes				Binary Outcomes			
	HIFM	EN	L	HL	HIFM	EN	L	HL
Pop 1:								
Median	**0.02**	0.37	0.35	0.19	**0.08**	11.27	2.54	0.99
Min	0.00	0.14	0.12	0.05	0.03	0.83	0.22	0.05
Max	0.36	15.60	1.23	1.57	0.21	102.11	20.84	8.26
Pop 2:								
Median	0.18	0.41	0.31	**0.17**	**0.13**	11.21	2.29	0.36
Min	0.06	0.17	0.08	0.07	0.06	0.71	0.16	0.11
Max	2.84	15.54	1.23	1.53	0.52	101.23	20.44	4.91

HIFM and Lasso. The simulation induces very strong sparsity, where the resulting coefficients are very close to zero, especially in the higher dimensional scenario. While Lasso may be overshrinking the signal in the data, which is why we see worse performance in prediction compared to elastic net and HIFM, the strong shrinkage results in better accuracy across all coefficients compared to elastic net. From these simulations HIFM shows that it is better at capturing both the coefficient estimates of the data and results in much improved prediction accuracy.

Lastly, we compare the number of factors used by HIFM for each population and compare those to the true number under simulation. Though we set $K = 10$ in simulation, we incorporate the weights in the loadings matrix that potentially shrink some of the factors across simulations. We set 0.05 as a threshold for considering whether that factor is included or not in the model when evaluating the number of factors chosen. This choice is arbitrary, but the results below were not sensitive to the chosen threshold within a reasonable range. For the HIFM we set $K = 5\log(p)$, where in this scenario $p = 50$ so k was set to 20 for the HIFM. We choose to look at the first example with 50 covariates for brevity. From Table 5 we see that, on average, HIFM selected around 11 factors for each population when all variables were continuous. For binary outcome (and half of all covariates being binary), HIFM selected 10 and 11 factors for first and second population, respectively. The true number of factors simulated averaged at around 10 factors for both populations and both types of outcomes, showing that our weighting mechanism was able to recover close to the truth. Figure 2 displays the resulting loadings matrix and the posterior mean of the weights post burn-in and thinning for both populations. The model selection properties using the weights are highlighted with the visualization, showing the shrinkage through the weights being used as a model selection tool for the number of factors to include in the model.

TABLE 5

Average number of factors selected by HIFM compared to truth with standard deviation in parentheses. Results displayed for simulations with $p = 50$ and $k = 10$, with HIFM k set to 20

	Pop. 1	Pop. 2
Continuous outcome:		
Normal HIFM	10.8 (1.2)	11.6 (2.1)
True	9.6 (0.6)	9.7 (0.7)
Binary outcome:		
Normal	10.2 (0.8)	11.1 (2.0)
True	9.9 (0.3)	9.8 (0.5)

For the 50 simulation runs when $p = 50$, we additionally tested three choices of $\alpha_0 = (5, 15, 50)$ to test how the setting of these hyperparameters may affect the predictive performance and the number of factors selected. We found the predictive performance is almost the exact same between the three choices (MSE = 0.13 for all choices of α_0 and AUC = 0.84 for $\alpha_0 = 5$ and 0.85 when $\alpha_0 = (15, 50)$), suggesting these hyperparameter choices do not have a strong effect on prediction. The three choices did result in different numbers of factors selected in the model: when $\alpha_0 = 5$, it selected 11 factors on average; when $\alpha_0 = 15$ and $\alpha_0 = 50$, the model selected 10 factors on average

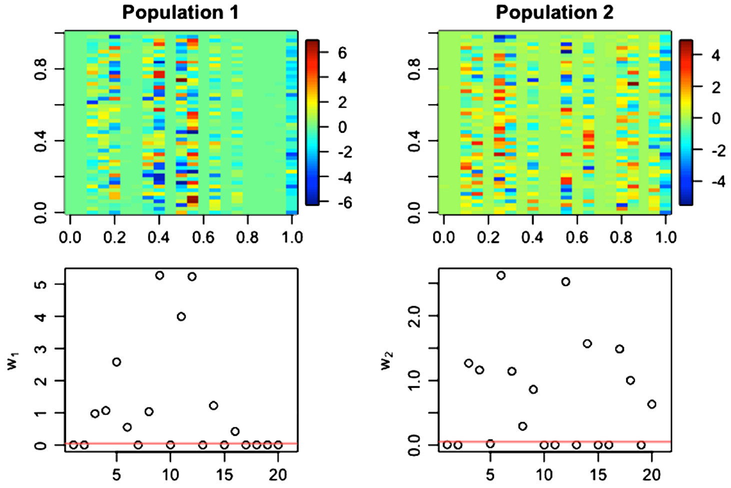

FIG. 2. *Visualization of loadings matrix for both simulated populations under HIFM learned with 20 factors. The image plot displays the posterior of the loadings matrix and the scatterplot displays the posterior mean of the weights, $\boldsymbol{w}_l$, where the red line indicates the chosen threshold used to determine number of factors in Table* 5.

4. Surgical complications.

4.1. *Goals, context and data*. The data in this experiment are derived from the repository, Pythia, of electronic health records (EHR) from all invasive surgical encounters from DUHS (Corey et al. (2018)). Invasive procedures are defined using the encounter's current procedural terminology (CPT) code and included all CPT codes that are identified by the Surgery Flag Software (AHRQ (2016)), and eliminated all patients under 18 years of age. Using data derived from the EHR provides the logistical benefit of easier implementation of the resulting tool in a clinical setting since the variables are conveniently found in a patient chart. However, EHR data are a by-product of day-to-day hospital activities, and the resulting data are known to be noisy and sparse. We therefore preprocessed the data to provide a cleaner and more manageable set of covariates to model.

We include covariates describing the surgical procedure, current medications of the patient, relevant comorbidities and other demographic information. The procedure information was captured by CPT codes and grouped into 128 procedure groupings categorized by the Clinical Classification Software (CCS). Procedural groupings with fewer than 200 total patients were removed and grouped into one larger miscellaneous category. This helped to assure that procedural effects were averaged across many patients and represented an overall effect size for geriatric patients and all surgical patients. We defined patient comorbidities by surveying all International Statistical Classification of Diseases (ICD) codes within one year preceding the date of the procedure and classified these diagnoses codes into 29 binary comorbidity groupings (S1) as defined by Elixhauser Comorbidity Index (Elixhauser et al. (1998)). We grouped the active outpatient medications recorded during medication reconciliation at preoperative visits into 15 therapeutic binary indicator features and created a separate feature that counted the total number of active medications. We define the outcomes, surgical complications, by diagnosis codes occurring within 30 days following the date of the invasive procedure. The outcomes were derived from 271 diagnosis codes and grouped into 12 categories that aligned with prior studies evaluating postsurgical complications (McDonald et al. (2018)) We use five of these outcomes to focus on the intervention goals of the POSH clinic. For example, neurological complications encompasses dementia, a common complication for patients over 65 and one that the POSH clinic specifically targets for their patients. The five outcomes modeled and reported below are cardiac complications, neurological complications, vascular complications, pulmonary complications and 90 day mortality. Mortality was identified as death occurring within 90 days of the index procedure date. Mortality is captured in the EHR during encounters for in-hospital death and uploaded from the Social Security Death Index for out-of-hospital deaths. Encounters missing EHR data were deemed not missing at random and were therefore excluded from the model development cohort. The resulting covariates are a mix of both continuous (BMI, age, etc.) and binary (indicator of comorbidities, etc), and therefore we utilize the probit transformation that was described above for all binary variables. In addition, we center and scale the continuous variables and also include an intercept in the model to learn the adjusted mean of the transformed binary variables.

We selected a cohort of 58,656 patients from Pythia that had undergone 77,150 invasive procedural encounters between January 2014 and January 2017, with all complete data. Of those encounters 22,055 are flagged as encounters that meet the POSH heuristic determined in clinical practice by surgeons and geriatricians, patient over the age of 85 OR a patient over the age of 65 with greater than five different medications, having two or more comorbidities, or whether the patient had a recent weight loss or signs of dementia. We form a binary variable to indicate whether a patient meets the POSH heuristic or not and use that grouping variable to determine the hierarchical structure in the factor model.

4.2. *Results*. Our interests are twofold: learn important subset of features and provide accurate predictions of risks of complication for both POSH and all surgical patients. Our goal is to show that pairing the POSH heuristic with a data-driven predictive modeling approach improves the triaging of patients into the high-risk clinic. Additionally, by understanding the covariates that most impact this high-risk geriatric population, we provide insights into the characteristics of the patient that make her/him high risk and therefore suggests other characteristics to be added to the current heuristic or develop possible interventions to target these characteristics.

We assess the performance of our model compared to other similar approaches. The first comparison approach is an infinite factor model (IFM), which follows the same Bayesian specification of the proposed HIFM but ignoring the grouping structure of the POSH group and the remaining surgery patients. The resulting model is essentially the exact model proposed but assumes all patients are from the same underlying population. Comparing the IFM to the HIFM will provide a sense of how important transfer learning or sharing information between the two populations improves prediction. The second comparison approach is a hierarchical factor model without any shrinkage. This follows the proposed model above without the use of the weighting mechanism derived from the HDP. Here, we get a sense of how a more common approach to factor modeling may work with multiple populations. One thing to note is that we still allow sharing in this model through the shared variance term in the each population's loadings matrix, ϕ_{jh}.

We trained each model on 60,000 encounters from the Pythia database and held out the 17,149 remaining encounters for validation, of which 4876 encounter met the POSH heuristic. We ran our Gibbs sampler for 4000 iterations, with burn-in of 2000 and thinned every six observations. The hyperparameters for the HDP were set to $\alpha_0 = 10$ and $\alpha_1 = \alpha_2 = 15$ with the tuning parameter for the Metropolis–Hastings step, $C = 50$. With 189 variables in our setting, we set the upper bound

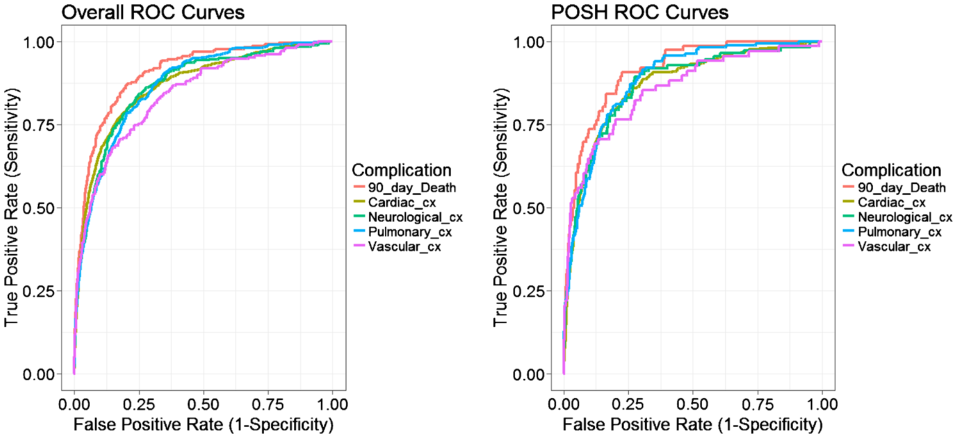

FIG. 3. *Receiver Operating Curve* (*ROC*) *of the five outcomes under the HIFM for encounters across the whole held-out test set and for the test set of geriatric patients. Posterior means with* 95% *credible intervals are displayed.*

for k equal to 30. We tested multiple different upper bounds and found this was sufficiently large where many of the factors were shrinking to zero.

To evaluate the predictive performance, we estimated the posterior predictive distribution and evaluated our predicted probabilities compared to the true outcomes. We use the posterior mean of the predictions and calculated the Receiver Operator Characteristic (ROC) curves for both the entire test set and then the POSH encounters within the test set. Figure 3 displays the resulting ROC curves from HIFM. All complications achieved strong performance with AUC between 0.84–0.91. Table 6 displays the resulting area under the ROC curves (AUC) and the area under precision-recall curves (AUPRC), comparing the overall test set and the POSH-only test set for the three different methods. For the outcomes, vascular and pulmonary complications, the IFM outperforms the HIFM for the overall population. However, HIFM outperforms IFM in the POSH group for these outcomes according to AUC. This suggests that the model ignoring the POSH group may help to improve learning overall for all patients but hinders the learning of the targeted geriatric group. In general, for the POSH patients HIFM outperforms the other two methods for all outcomes, according to AUC. There are a few AUPRC values that are higher in the IFM and HFM group, though those also correspond to AUCs that are lower than HIFM. These results suggest that our method is able to borrow strength from the larger group to improve the prediction for the smaller targeted group resulting in improved prediction, compared to the other two approaches tested.

In addition, we compare the sensitivities and specificities of the resulting model to those of the baseline POSH heuristic. Note that we remove the 500 patients that did go to the POSH clinic from the data so that we do not bias the results

TABLE 6

Classification results on five surgical outcomes, comparing full results and POSH specific results for the five outcomes under three different models. HIFM: hierarchical infinite factor model, IFM: infinite factor model, HFM: hierarchical factor model

	HIFM–Full		IFM–Full		HFM–Full	
	AUC	AUPRC	AUC	AUPRC	AUC	AUPRC
Mortality	**0.915**	0.130	0.913	**0.145**	0.913	0.137
Cardiac	**0.866**	**0.399**	0.856	0.383	0.858	0.394
Vascular	0.862	0.152	**0.864**	**0.171**	0.864	0.157
Neurological	**0.864**	**0.172**	0.860	0.159	0.857	0.144
Pulmonary	0.897	0.303	**0.898**	0.283	0.895	**0.307**
	HIFM–POSH		IFM–POSH		HFM–POSH	
	AUC	AUPRC	AUC	AUPRC	AUC	AUPRC
Mortality	**0.901**	0.187	0.873	**0.235**	0.867	0.187
Cardiac	**0.911**	0.209	0.842	**0.459**	0.806	0.432
Vascular	**0.867**	**0.402**	0.857	0.260	0.846	0.277
Neurological	**0.868**	**0.408**	0.842	0.205	0.836	0.216
Pulmonary	**0.872**	0.148	0.837	0.288	0.835	**0.302**

with possible treatment effects of the POSH clinics on the patients' outcomes. For the outcome death, the sensitivity and specificity for HIFM are 0.908 and 0.775, respectively. Alternatively, the POSH heuristic achieves a 0.345 sensitivity and 0.716 specificity. The POSH heuristic aims to target high risk patients, not necessarily defined to be high risk of death, though this outcome serves as the best proxy of overall risk. Currently, the POSH heuristic only identified 35% of patients that died, while using the HIFM model in conjunction with the heuristic improves sensitivity to 91%, providing evidence that our model is able to effectively identify those patients that are high risk and should go to POSH.

We next calculate the resulting coefficients derived for the POSH specific population from HIFM through the partitioned covariance matrix, discussed in Section 2.3, and find the posterior mean after burn-in and thinning. In Figure 4 we display the coefficients that are greater than 0.05 across all five outcomes along with their 95% credible intervals. Definitions for these variables are provided in Supplement C (Lorenzi, Henao and Heller (2019)). Different numbers of coefficients appear in each column of the plot that corresponds to each outcome, which is a result of the different levels of sparsity induced from the model. The resulting coefficients confirm existing knowledge in the literature of important covariates that predict these complications for geriatric patients. In addition, it suggests important procedures and medications that should be furthered flagged for patients to prevent higher risk of complications. Specifically, procedures for organ transplants, removal or insertion of a cardiac pacemaker and heart valve procedures increase

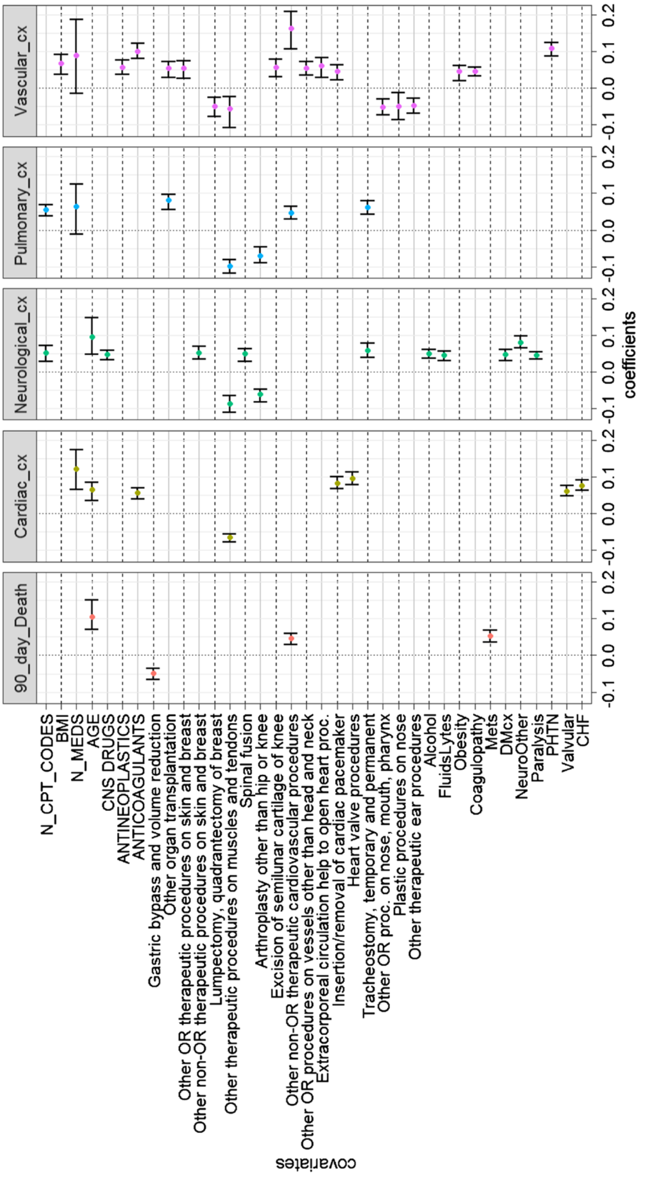

FIG. 4. *Largest estimated coefficients* ($|\beta| >= 0.05$) *for POSH group from HIFM. Posterior means with* 95% *credible intervals are plotted for each. See Supplement C for definitions of the variable names.*

the risk of cardiac complications. Some procedures are inherently less risky across the surgical outcomes, including procedures on muscles and tendons, joint replacements that are not hip or knee and procedures on the nose, mouth and ears. The number of medications patients take is strongly predictive of cardiac, pulmonary and vascular complications, and whether they are on anticoagulants increases the risk of vascular and cardiac complications. Risk factors for neurological complications, which includes dementia, are alcoholism, need for fluids and electrolytes (which indicates a nutritional deficiency), diabetes with complications, paralysis and previous neurological problems. These align well with the literature on risk factors of dementia, providing further evidence that our model detects predictive covariates that are specific to the geriatric population. In addition, an interesting feature of the chosen coefficients are their high correlation with one another. Typically in Lasso, highly correlated coefficients are shrunk so that only one remains in the model. A nice feature in our model is that we can characterize patients more accurately, regardless of how correlated the covariate space is, and provide a more accurate summary of important features. More importantly, these coefficients point to additional characteristics to better identify patients in the clinical setting.

5. Discussion.

We introduced the hierarchical infinite factor model that utilizes a hierarchical Dirichlet process weighting scheme as a sparsity-inducing transfer learning model. We contributed an easy-to-implement inference method and showed promising results that our method is effective at predicting surgical complications between unbalanced and sparse populations, in comparison to two other factor model approaches. Through simulation we show that, compared to state-of-the art baseline models, our model has better predictive accuracy and more accurate estimates of the coefficients, regardless of data size and type. In addition, simulations show that HIFM flexibly models each population with its own factor loadings matrix that controls the number of factors needed to best explain the data. The resulting factor scores are a new representation of the data that diminishes the distributional differences between the populations, resulting in similar predictive performance regardless if one population is smaller than the other.

Others in the literature have utilized transfer learning to improve prediction in health care settings. Gong et al. (2015) proposed an instance weighting algorithm used in risk stratification models of cardiac surgery using a weighting scheme based on distances of each observation to the mean of the target distribution's predictors. Wiens, Guttag and Horvitz (2014) discussed the problem of using data from multiple hospitals to predict hospital-associated infection with *Clostridium difficile* for a target hospital. Lee, Rubinfeld and Syed (2012) describe a method for transfer learning for the American College of Surgeon's National Surgical Quality Improvement Program (NSQIP) dataset, predicting mortality in patients after 30 days. Their methodology uses cost-sensitive support vector machines, first training the model on source data and next fitting the same model for the target data but regularizing the model parameters toward that of the source model. While these approaches succeed in accomplishing positive transfer in their individual applications, their methods fail to learn the dependence structure underlying the observed data and do not provide any uncertainty quantification to the predicted outcomes. Our approach not only achieves positive transfer learning such that prediction is improved in the target task, but it also provides interpretable insights into potential phenotypes of patients that best explain those at risk for complications post-surgery. We show above that using this predictive tool compared to the current POSH heuristic increases the sensitivity of death from 0.35 to 0.91. Improving sensitivity by almost a factor of 3 would have a huge impact for the geriatric patients at Duke. Implementing our proposed model in practice has the potential to save lives by either appropriately intervening on the patient or having further follow-up to decide whether the surgery is the right option for that patient.

While this work has focused on transfer learning between multiple populations, the model also shows promise as a sparsity inducing prior for single populations. In future work, we aim to develop this model further in two directions. First, as an improved transfer learning model that better shares information across multiple populations. With such large imbalances between geriatric and the full population and with low signal in many of the variables, the model often struggles to model the local population accurately, leading to more noise and less accurate predictions. In addition, the data contain many binary variables that require transformations to use with our model. Another avenue for future work is to better address this binary data type to reduce the additional uncertainty added to the inference through the mapping of the binary variables into the continuous space. The second direction will be to explore this model further as a sparse factor model, without explicitly aiming to perform transfer learning. The properties proved in Section 2.4 hold for a single population, therefore providing potential for further development as a shrinkage prior. Lastly, we look further to testing and evaluating this model on additional applications in the health realm. If the HIFM is applied to new types of data, new properties in the feature space, such as group-specific covariates or different data structures, will be of interest.

Additionally, one could consider the Laplace distribution, or commonly known as the double-exponential distribution, as a prior for the factors, $\boldsymbol{f}_i$. Laplace distributed factors provide two additional features to the model: First, it induces sparsity through the factor distribution, which may improve model fit in sparse settings. Second, it provides an improvement to the indeterminacy problems that occur naturally with Gaussian factor models. We studied our model with Laplace distributed factors and found that it provided no additional benefit in the prediction for our particular application, but in other settings, where identifiability is more of a concern, this is a reasonable alternative to the proposed model above.

Our work is a part of the continued effort to create a clinical platform to deliver individualized risk scores of complications at our university's health system for the purpose of triaging patients into preoperative clinics based on their underlying surgical risk. We plan to implement this framework directly into their electronic health system, so that clinicians will be able to assess the predicted complications directly through the patient's chart and treat the patient with suggested interventions that address the patient's increased risk.

Acknowledgments. The authors are grateful to Fan Li for providing helpful suggestions on the overall discussion of this work. We also acknowledge the Duke Institute for Health Innovations, specifically Kristin Corey, Sehj Kashyap, Mark Sendak and Suresh Balu for providing the data and clinical insights into the findings of our work. Lastly, thank you to the Department of Surgery at Duke University Health Systems for the opportunity to collaborate.

SUPPLEMENTARY MATERIAL

A. Proofs of HIFM properties (DOI: 10.1214/19-AOAS1292SUPPA; .pdf). Properties of hierarchical infinite factor model prior on loadings matrix.

B. Inference for full model (DOI: 10.1214/19-AOAS1292SUPPB; .pdf). All steps needed to sample the model.

C. Variable definitions shown in Figure 4. (DOI: 10.1214/19-AOAS1292SUPPC; .pdf). Description of variable names shown in Figure 4.

REFERENCES

AHRQ (2016). Healthcare cost and utilization project (hcup) surgery flag software. https://www.hcup-us.ahrq.gov/toolssoftware/surgflags/surgeryflags.jsp.

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

AVALOS-PACHECO, A., ROSSELL, D. and SAVAGE, R. S. (2018). Heterogeneous large datasets integration using Bayesian factor regression. Preprint. Available at arXiv:1810.09894.

BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. MR2806429

CARON, F. and DOUCET, A. (2008). Sparse Bayesian nonparametric regression. In *Proceedings of the* 25*th International Conference on Machine Learning* 88–95. ACM, New York.

CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008a). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456. MR2655722

CHEN, M., SILVA, J., PAISLEY, J., WANG, C., DUNSON, D. and CARIN, L. (2010). Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Trans. Signal Process.* **58** 6140–6155. MR2790088

COREY, K. M., KASHYAP, S., LORENZI, E., LAGOO-DEENADAYALAN, S. A., HELLER, K., WHALEN, K., BALU, S., HEFLIN, M. T., MCDONALD, S. R. et al. (2018). Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): A retrospective, single-site study. *PLoS Med.* **15** e1002701.

DESEBBE, O., LANZ, T., KAIN, Z. and CANNESSON, M. (2016). The perioperative surgical home: An innovative, patient-centred and cost-effective perioperative care model. *Anaesth. Crit. Care Pain Med.* **35** 59–66.

ELIXHAUSER, A., STEINER, C., HARRIS, D. R. and RM, C. (1998). Comorbidity measures for use with administrative data. *Med. Care* **36**.

ETZIONI, D. A., LIU, J. H., O'CONNELL, J. B., MAGGARD, M. A. and KO, C. Y. (2003). Elderly patients in surgical workloads: A population-based analysis. *Am. J. Surg.* **69** 961–965.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

GONG, J. J., SUNDT, T. M., RAWN, J. D. and GUTTAG, J. V. (2015). Instance weighting for patient-specific risk stratification models. In *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 369–378. ACM, New York.

HANOVER, N. (2001). Operative mortality with elective surgery in older adults. *Eff. Clin. Pract.* **4** 172–177.

HEALEY, M. A., SHACKFORD, S. R., OSLER, T. M., ROGERS, F. B. and BURNS, E. (2002). Complications in surgical patients. *Arch. Surg.* **137** 611–618.

ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. MR1952729

ISHWARAN, H. and JAMES, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information. *J. Comput. Graph. Statist.* **11** 508–532. MR1938445

JONES, T. S., DUNN, C. L., WU, D. S., CLEVELAND, J. C., KILE, D. and ROBINSON, T. N. (2013). Relationship between asking an older adult about falls and surgical outcomes. *J. Am. Med. Assoc. Surg.* **148** 1132–1138.

LEE, G., RUBINFELD, I. and SYED, Z. (2012). Adapting surgical models to individual hospitals using transfer learning. In *Data Mining Workshops* (*ICDMW*), 2012 *IEEE* 12*th International Conference on* 57–63. IEEE, New York.

LOPES, H. F. and WEST, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14** 41–67. MR2036762

LORENZI, E., HENAO, R. and HELLER, K. (2019). Supplement to "Hierarchical infinite factor models for improving the prediction of surgical complications for geriatric patients." DOI:10.1214/19-AOAS1292SUPPA, DOI:10.1214/19-AOAS1292SUPPB, DOI:10.1214/19-AOAS1292SUPPC.

LUCAS, J., CARVALHO, C., WANG, Q., BILD, A., NEVINS, J. and WEST, M. (2006). Sparse statistical modelling in gene expression genomics. *Bayesian Inference for Gene Expression and Proteomics* **1** 1.

MCDONALD, S. R., HEFLIN, M. T., WHITSON, H. E., DALTON, T. O., LIDSKY, M. E., LIU, P., POER, C. M., SLOANE, R., THACKER, J. K. et al. (2018). Association of integrated care coordination with postsurgical outcomes in high-risk older adults: The perioperative optimization of senior health (POSH) initiative. *JAMA Surg*. DOI: 10.1001/jamasurg.2017.5513

MCPARLAND, D., GORMLEY, I. C., MCCORMICK, T. H., CLARK, S. J., KABUDULA, C. W. and COLLINSON, M. A. (2014). Clustering South African households based on their asset status using latent variable models. *Ann. Appl. Stat.* **8** 747–776. MR3262533

MCPARLAND, D., PHILLIPS, C. M., BRENNAN, L., ROCHE, H. M. and GORMLEY, I. C. (2017). Clustering high-dimensional mixed data to uncover sub-phenotypes: Joint analysis of phenotypic and genotypic data. *Stat. Med.* **36** 4548–4569. MR3731239

MURPHY, K., GORMLEY, I. C. and VIROLI, C. (2017). Infinite mixtures of infinite factor analysers: Nonparametric model-based clustering via latent gaussian models. Preprint. Available at arXiv:1701.07010.

NI, Y., MUELLER, P. and JI, Y. (2018). Bayesian double feature allocation for phenotyping with electronic health records. Preprint. Available at arXiv:1809.08988.

POLSON, N. G. and SCOTT, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics* 9 501–538. Oxford Univ. Press, Oxford. MR3204017

RAVAL, M. V. and ESKANDARI, M. K. (2012). Outcomes of elective abdominal aortic aneurysm repair among the elderly: Endovascular versus open repair. *Surgery* **151** 245–260.

ROČKOVÁ, V. and GEORGE, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *J. Amer. Statist. Assoc.* **111** 1608–1622. MR3601721

SEO, D. M., GOLDSCHMIDT-CLERMONT, P. J. and WEST, M. (2007). Of mice and men: Space statistical modeling in cardiovascular genomics. *Ann. Appl. Stat.* **1** 152–178. MR2393845

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

SPEZIALE, G., NASSO, G., BARATTONI, M. C., ESPOSITO, G., POPOFF, G., ARGANO, V., GRECO, E., SCORCIN, M., ZUSSA, C. et al. (2011). Short-term and long-term results of cardiac surgery in elderly and very elderly patients. *J. Thorac. Cardiovasc. Surg.* **141** 725–731.

TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. MR2279480

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

WEST, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics*, 7 (*Tenerife*, 2002) 733–742. Oxford Univ. Press, New York. MR2003537

WIENS, J., GUTTAG, J. and HORVITZ, E. (2014). A study in transfer learning: Leveraging data from multiple hospitals to enhance hospital-specific predictions. *J. Am. Med. Inform. Assoc.* **21** 699–706.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

E. LORENZI
BERRY CONSULTANTS, LLC
3345 BEE CAVES RD, SUITE 201
AUSTIN, TEXAS 78746
USA
E-MAIL: elizabeth@berryconsultants.net

R. HENAO
DEPARTMENT OF BIOSTATISTICS
AND BIOINFORMATICS
DUKE UNIVERSITY
2424 ERWIN RD, HOCK PLAZA SUITE 1105
DURHAM, NORTH CAROLINA 27705
USA
E-MAIL: ricardo.henao@duke.edu

K. HELLER
DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
122 OLD CHEMISTRY BUILDING
DURHAM, NORTH CAROLINA 27708
USA
E-MAIL: katherine.heller@duke.edu