Published in final edited form as:

Biometrics. 2017 December ; 73(4): 1254–1265. doi:10.1111/biom.12690.

Analysis of Multiple Diverse Phenotypes via Semiparametric Canonical Correlation Analysis

Denis Agniel¹ and Tianxi Cai²

¹Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts

²Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts

Summary

Studying multiple outcomes simultaneously allows researchers to begin to identify underlying factors that affect all of a set of diseases (i.e., shared etiology) and what may give rise to differences in disorders between patients (i.e., disease subtypes). In this work, our goal is to build risk scores that are predictive of multiple phenotypes simultaneously and identify subpopulations at high risk of multiple phenotypes. Such analyses could yield insight into etiology or point to treatment and prevention strategies. The standard canonical correlation analysis (CCA) can be used to relate multiple continuous outcomes to multiple predictors. However, in order to capture the full complexity of a disorder, phenotypes may include a diverse range of data types, including binary, continuous, ordinal, and censored variables. When phenotypes are diverse in this way, standard CCA is not possible and no methods currently exist to model them jointly. In the presence of such complications, we propose a semi-parametric CCA method to develop risk scores that are predictive of multiple phenotypes. To guard against potential model mis-specification, we also propose a nonparametric calibration method to identify subgroups that are at high risk of multiple disorders. A resampling procedure is also developed to account for the variability in these estimates. Our method opens the door to synthesizing a wide array of data sources for the purposes of joint prediction.

Keywords

Canonical correlation analysis; Multiple phenotypes; Nonparametric calibration; Semi-parametric transformation models

1. Introduction

Studying complex disorders may require assessment of multiple outcomes for many possible reasons. A single complex disorder may be associated with multiple correlated phenotypes, or it may be multidimensional in nature and require multiple phenotypes to describe it adequately (Doshi-Velez, Ge and Kohane, 2014). Furthermore, some diseases may cross traditional diagnostic boundaries and require a diverse set of measurements to completely

Correspondence to: Denis Agniel.

Supplementary Materials: Web Appendices referenced in Sections 2.2, 4.2, and 4.2.1 are available with this article at the *Biometrics* website on Wiley Online Library.

characterize. In these cases, building prediction models for disease requires assessment of multiple phenotypes simultaneously. Similarly, multiple similar phenotypes may share etiological characteristics that can be discovered through a joint assessment (Alarcón-Segovia, 2005).

The goal of this article is to identify patient characteristics that are predictive of multiple phenotypes and identify subgroups at high risk of multiple disorders. Suppose, $\mathbf{x} = (x_j)_{1 \leq j \leq p}$ is a set of candidate genomic markers or risk factors for a set of phenotypes $\mathbf{y} = (y_m)_{1 \leq m \leq M}$. In order to tailor treatment strategies or prevention mechanisms to patients, it may be informative to produce a *risk score* based on $\mathbf{x}$ that differentiates patients phenotypically, where, for example, patients with high values of the risk score are at high risk for many diseases and patients with low values are at low risk. One way to build this type of risk score is to relate it to a *trait* that combines information across all phenotypes. In many situations, traits are not available *a priori* and must be derived from the data.

One classical approach to build such prediction models is canonical correlation analysis (CCA) (Hotelling, 1936; Hardoon, Szedmak and Shawe-Taylor, 2004). In CCA, traits and risk scores are identified as linear combinations $\mathbf{a}_1^\top \mathbf{y}$ and $\mathbf{b}_1^\top \mathbf{x}$, respectively, via the maximization problem

$$\operatorname{argmax}_{\mathbf{a}, \mathbf{b}} \operatorname{cor}(\mathbf{a}^\top \mathbf{y}, \mathbf{b}^\top \mathbf{x}), \quad (1)$$

subject to a constraint on the norms of $\mathbf{a}$ and $\mathbf{b}$ for identifiability. Additional projections pairs $(\mathbf{a}_k, \mathbf{b}_k)_{k \geq 1}$, orthogonal to $(\mathbf{a}_1, \mathbf{b}_1)$ and each other, can also be sequentially identified to capture residual correlations between the outcomes and predictors. Various extensions of CCA have been proposed in recent years. Other measures of association beyond Pearson's correlation coefficient (Jin and Cui, 2010) and semi-parametric single-index models (Xia, 2008) have been proposed to account for nonlinearity between the traits and risk profiles. Kernel methods have also been proposed in Hardoon, Szedmak and Shawe-Taylor (2004). Zhu, Zhu and Li (2007) estimated transformations of $\mathbf{y}$ via smoothing and perform partial least squares, a close relative of CCA. The Gifi system (Michailidis and de Leeuw, 1998) includes many methods for nonlinear multivariate analysis of categorical $\mathbf{y}$, including methods for a version of nonlinear CCA by minimizing the sum of squares between a latent canonical variate and linear combinations of indicators for all values of $\mathbf{x}$ and $\mathbf{y}$.

However, these methods are inappropriate when $\mathbf{y}$ are not fully observed due to censoring or truncation or may contain a diverse set of measurements representing different data types. For example, $\mathbf{y}$ may include binary (such as diagnosis of disease), continuous (like a biomarker for disease), ordinal (severity of disease), and censored variables (time until onset of symptoms). We propose semi-parametric CCA (sCCA), which allows us to synthesize the diverse information in $\mathbf{y}$ and build joint prediction models for all phenotypes simultaneously. The sCCA approach identifies projection pairs $(\mathbf{a}_k, \mathbf{b}_k)_{k \geq 1}$ by solving an equation like (1), but the trait $\mathbf{a}^\top \mathbf{y}$ is replaced by $\mathbf{a}^\top \mathbf{h}$ where $\mathbf{h}$ is a vector of transformed (possibly latent)

phenotypes, all of which are on the same scale. Through this transformed set of phenotypes, we may extend the existing machinery of CCA to produce risk scores that can optimally predict meaningful latent traits related to multiple phenotypes.

In this article, we obtain such transformed outcomes $\mathbf{h}$ via a normal semi-parametric transformation model (nSTM) framework (Zeng and Lin, 2010; Lin et al., 2013). Few methods exist to jointly model $\mathbf{y}$ that are subject to censoring or truncation. Vector generalized linear models (VGLMs) have been proposed to model multiple exponential family random variables via copulas (Song, Li and Yuan, 2009). These models have been generalized to a slightly larger class of multivariate parametric likelihoods that can be solved using weighted least squares (Yee, 2015), but the VGLM methods are not applicable to censored or truncated data. Estimation procedures have also been proposed for nSTM to model multiple phenotypes that are potentially of mixed types and/or subject to censoring (Othus and Li, 2010; Snaveley, Harrington and Li, 2014; Zhou et al., 2014). These previous approaches to nSTM have featured restrictive assumptions about the covariance structure and cumbersome iterative methods for estimation while lacking asymptotic results for censored outcomes. Furthermore, existing literature on VGLM and nSTM largely focuses on estimation of the regression model parameters but cannot be directly used to derive risk profiles that are predictive of latent traits. Here, we propose the use of nSTM in the context of CCA to derive such risk profiles and also provide a robust procedure to identify subpopulations that are at high risk of multiple phenotypes via an additional nonparametric calibration step.

The first main contribution of this article is to provide risk prediction methods in an inherently challenging setting – multiple outcomes of diverse types, possibly subject to censoring. We employ the nSTM framework and provide general yet relatively simple inference procedures for the sCCA directions under this setting without requiring restrictive assumptions about the covariance structure. The second main contribution of this article is to provide robust methods for individualized simultaneous prediction of multiple phenotypes via nonparametric calibration, recognizing that the models used to build risk scores may not always be correctly specified. This two-step approach, sCCA followed by nonparametric calibration, uses the sCCA framework to group patients according to univariate risk scores that capture different underlying phenotypic traits. Subsequently via calibration, we identify subgroups at high risk of multiple phenotypes without relying on the correct specification of the semi-parametric models employed in the sCCA step.

The structure of the rest of the article is as follows. In Section 2, we introduce sCCA, and discuss its properties. In Section 3, we lay out the nonparametric calibration method for simultaneous prediction of multiple phenotypes and identifying high-risk subgroups. In Section 4, we apply sCCA to a genetic study to identify risk profiles for autoimmune disease, and we provide simulation results. Finally, in Section 5, we provide a discussion of the method.

2. Semiparametric Canonical Correlation Analysis

Suppose the data for analysis consists of n independent and identically distributed (iid) random vectors $\{\mathbf{V}_i = (\mathbf{y}_i^\top, \boldsymbol{\delta}_i^\top, \mathbf{x}_i^\top)^\top\}_{1 \leq i \leq n}$ where $\mathbf{y}_i = (y_{im})_{1 \leq m \leq M}$ is a set of M phenotypes that may be of diverse scale and subject to censoring, $\boldsymbol{\delta}_i = (\delta_{im})_{1 \leq m \leq M}$, δ_{im} is an indicator that y_{im} is uncensored, and $\mathbf{x}_i = (x_{ij})_{1 \leq j \leq p}$ is a set of p predictors for the i th subject. We assume without loss of generality that x_{ij} has been standardized such that $E(x_{ij}) = 0$, $\text{var}(x_{ij}) = 1$. When y_{im} is subject to censoring, we assume that $y_{im} = \min(z_{im}, C_{im})$, C_{im} is the censoring time independent of the underlying survival time z_{im} and $\delta_{im} = \mathbb{I}(z_{im} < C_{im})$. When y_{im} is not subject to censoring, we let $\delta_{im} = 1$.

Recall that, if we were interested in forming optimal linear scores for predicting traits of the form $\mathbf{a}_k^\top \mathbf{y}$, we would seek to obtain (1), or equivalently, $\arg\max_{\mathbf{a}, \mathbf{b}} \mathbf{a}^\top \Sigma_{yx} \mathbf{b}$ subject to $\mathbf{a}^\top \Sigma_{yy} \mathbf{a} = \mathbf{b}^\top \Sigma_{xx} \mathbf{b} = 1$ where $\Sigma_{wv} = \text{cov}(\mathbf{w}, \mathbf{v})$. However, due to the components of $\mathbf{y}$ being incompletely observed or measured on incomparable scales, identifying traits as linear combinations of $\mathbf{y}$ may not be possible or clinically meaningful. Instead, we consider traits to be of the form $\mathbf{a}_k^\top \mathbf{h}$ and we thus obtain

$$(\mathbf{a}_1, \mathbf{b}_1) = \arg\max_{\mathbf{a}, \mathbf{b}} \mathbf{a}^\top \Sigma_{h,x} \mathbf{b}, \quad \text{subject to } \mathbf{a}^\top \Sigma_{hh} \mathbf{a} = \mathbf{b}^\top \Sigma_{xx} \mathbf{b} = 1, \quad (2)$$

where $\mathbf{h}$ is a vector of transformed phenotypes all on similar scales, the exact form of which will be given in equation (4) of Section 2.1. Additional directions $(\mathbf{a}_k, \mathbf{b}_k)$ can be obtained by solving (2) with the additional constraints that $(\mathbf{a}_k^\top \mathbf{a}_j = \mathbf{b}_k^\top \mathbf{b}_j = 0)_{1 \leq j < k}$. One may summarize the predictiveness of $\mathbf{b}_k^\top \mathbf{x}$ for the latent trait $\mathbf{a}_k^\top \mathbf{h}$ based on the *proportion of variance explained* $r_k = \text{tr}(\sum_{hh} \mathbf{a}_k \rho_k^2 \mathbf{a}_k^\top \sum_{hh}) / \text{tr}(\sum_{hh})$ (Glahn, 1968), where $\rho_k = \text{cor}(\mathbf{a}_k^\top \mathbf{h}, \mathbf{b}_k^\top \mathbf{x})$.

The parameters of interest, $\boldsymbol{\theta} = (\mathbf{a}_k, \mathbf{b}_k, r_k)_{1 \leq k \leq K}$ can be expressed as $\boldsymbol{\theta} = \mathcal{F}(\boldsymbol{\sigma})$, where $\boldsymbol{\sigma} = (\text{vec}(\Sigma_{xx})^\top, \text{vec}(\Sigma_{hx})^\top, \text{vec}(\Sigma_{hh})^\top)^\top$ and $\mathcal{F}(\cdot)$ is some known smooth function. Thus, once $\boldsymbol{\sigma}$ is estimated by a consistent estimator $\hat{\boldsymbol{\sigma}}$, a plug-in estimator of $\boldsymbol{\theta}$ may be constructed as $\hat{\boldsymbol{\theta}} = \mathcal{F}(\hat{\boldsymbol{\sigma}})$. We next identify $\mathbf{h}$, Σ_{hh} and Σ_{hx} through latent variable modeling via the nSTM framework in Section 2.1.

Remark on Using $\mathbf{h}$. We use transformed latent phenotypes $\mathbf{h}$ to form an underlying trait $\mathbf{a}^\top \mathbf{h}$ for several reasons. First, the observed outcomes are of different scales and may consist of binary, censored, and continuous outcomes of drastically different scales. As such, a linear combinations of the observed $\mathbf{y}$ may not be clinically meaningful. Even for continuous outcomes, it is common to choose a transformation of a phenotype such that its relationship to predictors is closer to linearity. It is, for example common to transform viral load in HIV research (O'Brien et al., 1996). Second, when analyzing binary outcomes, latent thresholding models are often used to interpret the observed binary outcome as a thresholded version of a latent continuous phenotype. Similarly, when analyzing survival outcomes, the observed data consist of y , the minimum of survival time z and the censoring time C , while the

phenotype of interest is obviously the latent event time z . Combining such a y with other outcomes would not be clinically meaningful. Under the proposed semi-parametric transformation models, the transformed latent phenotypes, $\mathbf{h}$, all relate to the covariates $\mathbf{x}$ linearly and the corresponding residuals are of similar scales, which makes it more meaningful to linearly combine $\mathbf{h}$.

2.1. Joint Modeling of Multiple Phenotypes via nSTM

To relate $\mathbf{y}$ of diverse scales to $\mathbf{x}$ in a comparable fashion, we assume that the observed $\mathbf{y}$ represents potentially partially observed versions of latent variables $\mathbf{z} = (z_m)_{1 \leq m \leq M}$ whose components can be put on comparable scales after appropriate transformations. The relationship between the observed y_m and the latent z_m can be described as follows. If y_m is continuous and not subject to censoring, then $y_m = z_m$ and $\delta_m = 1$. When y_m is subject to censoring, then $y_m = \min(z_m, C_m)$ and $\delta_m = I\{z_m \leq C_m\}$. For a discrete y_m , without loss of generality, we assume that y_m takes values in $\mathcal{Y}_m = \{0, 1, \dots, K_m\}$. We consider y_m to be a thresholded version of z_m such that

$y_m = \sum_{k=1}^{K_m-1} k \times I\{z_m \in (k-1, k]\} + K_m \times I\{z_m > K_{m-1}\}$. For $m = 1, \dots, M$, we first assume that the marginal distribution of z_m follows the following STM

$$P(z_m \leq z | \mathbf{x}) = \Phi\{\mathbf{x}^\top \boldsymbol{\beta}_m + h_m(z)\}, \quad (3)$$

where $\boldsymbol{\beta}_m$ represents the unknown effect of $\mathbf{x}$ on z_m and the link function Φ is the standard normal distribution function. The transformation $h_m(\cdot)$ is an infinite-dimensional, unspecified, smooth, increasing function if y_m is continuous, while it is only identifiable on the set $\{0, \dots, K_m - 1\}$ if y_m is discrete. If y_m were categorical and unordered with K_m levels, it may be modeled as $K_m - 1$ binary variables via various multinomial regression models.

The model (3) is equivalent to

$$h_m(z_m) = -\mathbf{x}^\top \boldsymbol{\beta}_m + \varepsilon_m, \quad \varepsilon_m \sim N(0, 1), \quad (4)$$

and we thus obtain $\mathbf{h} = (h_m(z_m))_{1 \leq m \leq M}$, the $M \times 1$ vector of transformed latent phenotypes used in (2). For context, if ε_m were instead taken to be from the extreme value distribution, (4) would correspond to a proportional hazards model. In this formulation, the components of $\mathbf{h}$ are on comparable scales, as each is expressed as a linear combination of $\mathbf{x}$ and an independent standard normal error. The specification of the joint distribution of the data follows naturally from the marginal models (4). Taking $\boldsymbol{\varepsilon} = (\varepsilon_m)_{1 \leq m \leq M}$ to be jointly normal and recognizing that $\boldsymbol{\varepsilon}_m = \mathbf{x}^\top \boldsymbol{\beta}_m + h_m(z_m)$, the distribution for $\mathbf{z}$ conditional on $\mathbf{x}$ may be written

$$F_{\mathbf{x}}(\mathbf{z}) = \Phi_{\sum_{\varepsilon\varepsilon}} \{ \mathbf{x}^\top \boldsymbol{\beta}_1 + h_1(z_1), \dots, \mathbf{x}^\top \boldsymbol{\beta}_M + h_M(z_M) \}, \text{ where } \sum_{\varepsilon\varepsilon} = \text{cor}(\varepsilon, \varepsilon). \quad (5)$$

Throughout, $\Phi_{\sum_{\varepsilon\varepsilon}}(\cdot)$ and $\phi_{\sum_{\varepsilon\varepsilon}}$ denote the distribution and density functions of $N(\mathbf{0}, \sum_{\varepsilon\varepsilon})$, respectively.

Since $\mathbf{z}$ is not always observed, the likelihood function for model parameters given the observed data involves integrating over the region of possible values that $\mathbf{z}$ may take given $(\mathbf{y}, \boldsymbol{\delta})$, denoted by $\mathcal{A}(\mathbf{y}, \boldsymbol{\delta})$. The likelihood for $\mathbb{V}$ can be written

$$\mathcal{L}(\sum_{\varepsilon\varepsilon}, \mathcal{H}, ;) = \prod_{i=1}^n \int_{\mathcal{A}(\mathbf{y}_i, \boldsymbol{\delta}_i)} \phi_{\sum_{\varepsilon\varepsilon}} \{ \mathbf{x}_i^\top \boldsymbol{\beta}_1 + h_1(t_1), \dots, \mathbf{x}_i^\top \boldsymbol{\beta}_M + h_M(t_M) \} dt$$

where $\mathbb{B}$ is a $p \times M$ matrix with the m th column being $\boldsymbol{\beta}_m$, $\phi_{\sum_{\varepsilon\varepsilon}}$ is the multivariate normal density with mean 0 and covariance matrix $\sum_{\varepsilon\varepsilon}$, $\mathbf{t} = (t_m)_{m=1}^M$ and $\mathcal{H} = (h_m(\cdot))_{m=1}^M$ is the set of unknown transformation functions. To provide more intuition on the likelihood representation for different outcomes and for the implied region $\mathcal{A}(\mathbf{y}, \boldsymbol{\delta})$, consider the following examples with $M=2$ and various types of outcomes. If $\mathbf{z}$ is fully observed with $\mathbf{z} = \mathbf{y}$, then

$$\mathcal{L}(\sum_{\varepsilon\varepsilon}, \mathbf{h}, ;) = \prod_{i=1}^n \phi_{\sum_{\varepsilon\varepsilon}} \{ \mathbf{x}_i^\top \boldsymbol{\beta}_1 + h_1(y_{i1}), \mathbf{x}_i^\top \boldsymbol{\beta}_2 + h_2(y_{i2}) \}.$$

If y_1 is binary and y_2 is potentially right-censored, then

$$\begin{aligned} & \mathcal{L}(\sum_{\varepsilon\varepsilon}, \mathbf{h}, ;) \\ &= \prod_{i=1}^n \left(\int_{-\infty}^0 \phi_{\sum_{\varepsilon\varepsilon}} \{ \mathbf{x}_i^\top \boldsymbol{\beta}_1 + h_1(t), \mathbf{x}_i^\top \boldsymbol{\beta}_2 + h_2(y_{i2}) \} dt \right)^{\delta_{i2}(1-y_{i1})} \\ & \quad \times \left(\int_0^\infty \phi_{\sum_{\varepsilon\varepsilon}} \{ \mathbf{x}_i^\top \boldsymbol{\beta}_1 + h_1(t), \mathbf{x}_i^\top \boldsymbol{\beta}_2 + h_2(y_{i2}) \} dt \right)^{\delta_{i2}y_{i1}} \\ & \times \left(\int_{y_{i2}}^\infty \int_{-\infty}^0 \phi_{\sum_{\varepsilon\varepsilon}} \{ \mathbf{x}_i^\top \boldsymbol{\beta}_1 + h_1(t_1), \mathbf{x}_i^\top \boldsymbol{\beta}_2 + h_2(t_2) \} dt_1 dt_2 \right)^{(1-\delta_{i2})(1-y_{i1})} \\ & \quad \times \left(\int_{y_{i2}}^\infty \int_0^\infty \phi_{\sum_{\varepsilon\varepsilon}} \{ \mathbf{x}_i^\top \boldsymbol{\beta}_1 + h_1(t_1), \mathbf{x}_i^\top \boldsymbol{\beta}_2 + h_2(t_2) \} dt_1 dt_2 \right)^{(1-\delta_{i2})y_{i1}} \end{aligned}$$

2.2. Estimating σ

In this section, we propose estimators for quantities relevant to sCCA, namely $\sum_{xx}$, $\sum_{hx}$, and $\sum_{hh}$. Or, equivalently, we propose estimators for $\sum_{xx}$, $\boldsymbol{\beta}_m$, and $\sum_{\varepsilon\varepsilon}$, since $\sum_{hx}$ and $\sum_{hh}$ may be written as

$$\sum_{hh} = \text{cov}(\mathbf{h}, \mathbf{h}) = \sum_{xx} + \sum_{\varepsilon\varepsilon} \text{ and } \sum_{hx} = \text{cov}(\mathbf{h}, \mathbf{x}) = \sum_{xx}. \quad (6)$$

Estimation of Σ_{xx} may proceed in the usual fashion, using the standard unbiased estimator

$$\hat{\Sigma}_{xx} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top, \quad (7)$$

Next, we may estimate $\mathbb{B}$ and $\mathcal{H}$ respectively as $\hat{\mathbb{B}} = [\hat{\boldsymbol{\beta}}_1, \dots, \hat{\boldsymbol{\beta}}_M]$ and $\hat{\mathcal{H}} = (\hat{h}_m(\cdot))_{1 \leq m \leq M}$ where $\{\hat{\boldsymbol{\beta}}_m, \hat{h}_m(\cdot)\}$ is the marginal nonparametric maximum likelihood estimate of $\{\boldsymbol{\beta}_m, h_m(\cdot)\}$ (NPMLE) by maximizing each marginal likelihood specified by (4) individually (Murphy and Van der Vaart, 2000). The parameters are directly analogous to regression parameters ($\boldsymbol{\beta}_m$) and baseline hazard (h_m) in a proportional hazards model, and likelihood maximization may proceed using standard methods for fitting a proportional odds model with probit link (Venables and Ripley, 2002; Agresti and Kateri, 2011). With $\hat{\mathbb{B}}$ and $\hat{\mathcal{H}}$, we subsequently estimate Σ_{ee} , as the maximizer of the pseudo-likelihood derived from (6),

$$\hat{\Sigma}_{ee} = \underset{\Sigma_{ee}}{\operatorname{argmax}} \mathcal{L}(\Sigma_{ee}, \hat{\mathcal{H}}; \hat{\mathbb{B}}). \quad (8)$$

Final estimates of $\boldsymbol{\sigma}$ are obtained from (7) and by plugging in (8), (7), and $\hat{\mathbb{B}}$ into (6).

Remark 1. The two stage estimation – by first maximizing the marginal likelihoods (4) and then estimating the correlation structure by maximizing the pseudo-likelihood as in (8) – is appealing for its simplicity. Under semi-parametric settings, directly maximizing the joint likelihood with respect to multiple finite- and infinite-dimensional parameters are both computationally and inferentially challenging, particularly when the outcomes are subject to truncation and censoring. This type of two-stage approach, suggested as generally efficient, is commonly used in the copula literature (Joe, 2005).

2.3. Assessing Variability and Correcting Bias

To make inference about $\hat{\boldsymbol{\theta}} = \mathcal{F}(\hat{\boldsymbol{\sigma}})$, we show in web appendix A that $n^{\frac{1}{2}}(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma})$ has an influence function expansion of $n^{-\frac{1}{2}} \sum_{i=1}^n \mathcal{U}_{\sigma i}(\boldsymbol{\sigma})$. This, together with a Taylor series expansion,

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = n^{\frac{1}{2}}\{\mathcal{F}(\hat{\boldsymbol{\sigma}}) - \mathcal{F}(\boldsymbol{\sigma})\} = n^{\frac{1}{2}}\dot{\mathcal{F}}(\boldsymbol{\sigma})(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}) + n^{\frac{1}{2}}O_p(\|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|^2),$$

implies that $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow N(\mathbf{0}, \mathbf{C})$ in distribution, where $\mathbf{C} = \dot{\mathcal{F}}(\boldsymbol{\sigma}) \operatorname{var}\{\mathcal{U}_{\sigma i}(\boldsymbol{\sigma})\} \dot{\mathcal{F}}(\boldsymbol{\sigma})^\top$, and $\dot{\mathcal{F}}(\boldsymbol{\sigma}) = \dot{\mathcal{F}}(\boldsymbol{\sigma}) / \boldsymbol{\sigma}^\top$ whose explicit forms can be found in Ogasawara (2007). However, the variance estimates based on the asymptotic result may be unstable in finite samples, as noted in Ogasawara (2007) even in a traditional CCA setting. Furthermore, explicit estimation of $\mathcal{U}_{\sigma i}(\boldsymbol{\sigma})$ is generally difficult in practice. Instead, we propose to estimate $\mathbf{C}$ via perturbation-resampling (Jin, Ying and Wei, 2001; Cai, Tian and Wei, 2005).

Let $\Xi = (\xi_i)_{i=1}^n$ be iid non-negative random variables with $E(\xi_i) = 1$ and $\text{var}(\xi_i) = 1$, generated independently of $\mathbb{V}$. We obtain $\hat{\boldsymbol{\sigma}}^*$ by weighting each individual by ξ_i in (4), (7), and (8); and let $\hat{\boldsymbol{\theta}}^* = \mathcal{F}(\hat{\boldsymbol{\sigma}}^*)$. Generating $\mathcal{Q}$ independent sets of Ξ , we obtain $\mathcal{Q}$ realizations of $\hat{\boldsymbol{\theta}}^*$, denoted by $\{\hat{\boldsymbol{\theta}}^{(i)}\}_{i=1}^{\mathcal{Q}}$. The empirical distribution of $\{n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^{(i)} - \hat{\boldsymbol{\theta}})\}_{1 \leq i \leq \mathcal{Q}}$ conditional on $\mathbb{V}$ can be used to approximate the distribution of $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$. One may construct a $(1 - \alpha)$ % confidence interval (CI) for the j th component of $\boldsymbol{\theta}$, θ_j based on the empirical standard error or quantiles of $\{\hat{\theta}_j^{(i)}\}_{1 \leq i \leq \mathcal{Q}}$. In practice, we find that the quantile-based CIs perform better due to the skewness of the distributions in finite samples. The bootstrap, which may be construed as drawing the full vector Ξ from a multinomial distribution, may also be used equivalently.

Although $\hat{\boldsymbol{\theta}}$ is asymptotically unbiased, in small samples some of its components might be biased due to overfitting, particularly $\hat{r}_k$. One may use the perturbed realizations to estimate the bias and produce a corrected estimator. One may let $\hat{\mathcal{B}} = \mathcal{Q}^{-1} \sum_{i=1}^{\mathcal{Q}} (\hat{\boldsymbol{\theta}}^{(i)} - \boldsymbol{\theta})$ be an estimator of $E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, and let the bias-adjusted estimator be denoted $\hat{\boldsymbol{\theta}}_{\text{BC}} = \hat{\boldsymbol{\theta}} - \hat{\mathcal{B}}$.

3. Individualized Prediction of Multiple Phenotypes via Nonparametric Calibration

The risk score $\hat{\mathbf{b}}_k^\top \mathbf{x}$ is expected to be highly predictive of the latent trait $\hat{\mathbf{a}}_k^\top \mathbf{h}$ and $\hat{r}_k$ summarizes the predictiveness of $\hat{\mathbf{b}}_k^\top \mathbf{x}$ for this latent trait under the correct specification of the STMs. However, under model mis-specification, $\hat{r}_k$ may be a biased estimate of the true prediction performance. In addition, it would be of great interest to assess the predictiveness of $\hat{\mathbf{b}}_k^\top \mathbf{x}$ for each of the individual observed phenotypes and identify subgroups of patients who may be at high risk of developing several phenotypes. This motivates us to derive nonparametrically calibrated estimates of the subgroup-specific risk for each phenotype of interest.

Let $\bar{\mathbf{b}}_k$ denote the limit of $\hat{\mathbf{b}}_k$ regardless of model adequacy. For any function $f(\cdot)$, we may estimate $E\{f(z_{im}) | \bar{\mathbf{b}}_k^\top \mathbf{x}_i = x\}$ nonparametrically via kernel smoothing as

$$\hat{\mathbb{E}}_f(x, \bar{\mathbf{b}}_k) = \frac{n^{-1} \sum_{i=1}^n K_h\{\Phi(x) - \Phi(\hat{\mathbf{b}}_k^\top \mathbf{x}_i)\} f(y_{im})}{n^{-1} \sum_{i=1}^n K_h\{\Phi(x) - \Phi(\hat{\mathbf{b}}_k^\top \mathbf{x}_i)\}} \quad (9)$$

if y_{im} is uncensored, where $K_h(u) = h^{-1}K(u/h)$ with kernel function $K(\cdot)$ and the transformation $\Phi(\cdot)$ is used to improve the kernel smoothing similar to those suggested in Wand, Marron and Ruppert (1991). Various choices of $f(\cdot)$ might be of interest. Examples include the conditional trimmed mean, $\mu_{km}(x) = E\{\mathbb{I}_{\tau_1, \tau_2}(z_{im}) | \bar{\mathbf{b}}_k^\top \mathbf{x}_i = x\}$ as $\hat{\mu}_{km}(x) = \hat{\mathbb{E}}_{\mathbb{I}_{\tau_1, \tau_2}}(x, \hat{\mathbf{b}})$, and the conditional distribution function $\mathbb{F}_{km}(c|x) = P(z_{im} \leq c | \bar{\mathbf{b}}_k^\top \mathbf{x}_i = x)$ as $\hat{\mathbb{F}}_{km}(c|x)$

$= \hat{\mathbb{F}}_{I_c}(x, \hat{\mathbf{b}}_k)$, where $\mathbb{I}_{\tau_1, \tau_2}(z) = \max\{\min(z, \tau_2), \tau_1\}$ and $I_c(y_{im}) = \mathbb{I}(y_{im} - c)$. Here and in the sequel, for notational ease, the dependence of $\hat{\mu}_{km}(x)$ and $\mu_{km}(x)$ on τ_1 and τ_2 will be suppressed. For uncensored data, we assume that $\mathbb{I}(y_{im} - c) = \mathbb{I}(z_{im} - c)$ and $\mathbb{I}_{\tau_1, \tau_2}(z_{im}) = \mathbb{I}_{\tau_1, \tau_2}(y_{im})$ so that these quantities are observable.

For censored outcomes, conditional Nelson–Aalen or Kaplan–Meier estimators may be used to estimate the conditional survival function given the risk score (Cai et al., 2010), and the methods in Tian, Zhao and Wei (2014) may be used to estimate the conditional trimmed mean. The appropriate choice of the bandwidth h for each k , denoted by h_k , can be selected individually via the standard cross-validation.

In order to evaluate what regions of $\mathbf{x}$ may indicate high risk of multiple disorders, one may select clinically meaningful cut-off values $\mathbf{c}^\top = (c_m)_{1 \leq m \leq M}$ for the M phenotypes and simultaneously estimate $\mathbb{F}_{km}(c_m | \mathbf{x})$. Here, we assume without loss of generality that phenotypes are defined in such a way that higher values of z_m correspond to higher risk of disorder. For each subset of phenotypes $\mathcal{M} \subset \{1, \dots, M\}$ and a set of desired risk threshold levels $\boldsymbol{\omega}^\top = (\omega_m)_{1 \leq m \leq M}$, we may identify the high-risk subpopulation

$\Omega_{\mathcal{M}, k}(\boldsymbol{\omega}) = \{\mathbf{x} : \mathbb{F}_{km}(c_m | \bar{\mathbf{b}}_k^\top \mathbf{x}) \leq \omega_m, m \in \mathcal{M}\}$, based on the risk score $\hat{\mathbf{b}}_k^\top \mathbf{x}$ using the

nonparametric estimate $\hat{\Omega}_{\mathcal{M}, k}(\boldsymbol{\omega}) = \{\mathbf{x} : \hat{\mathbb{F}}_{km}(c_m | \hat{\mathbf{b}}_k^\top \mathbf{x}) \geq \omega_m, m \in \mathcal{M}\}$. We note that for

different choices of k , $\mathcal{M}$ may be chosen accordingly since the risk score $\hat{\mathbf{b}}_k^\top \mathbf{x}$ may only be predictive of a subset of the phenotypes. Alternatively, an overall assessment of risk can be based on the conditional trimmed mean. High-risk subgroups can then be identified as

$\Omega_{\mathcal{M}, k}(\boldsymbol{\mu}; \boldsymbol{\omega}) = \{\mathbf{x} : \mu_{km}(\bar{\mathbf{b}}_k^\top \mathbf{x}) \geq \omega_m, m \in \mathcal{M}\}$, which can be consistently estimated as

$\hat{\Omega}_{\mathcal{M}, k}(\hat{\boldsymbol{\mu}}; \boldsymbol{\omega}) = \{\mathbf{x} : \hat{\mu}_{km}(\hat{\mathbf{b}}_k^\top \mathbf{x}) \geq \omega_m, m \in \mathcal{M}\}$.

To account for sampling variation, we now turn attention to providing confidence sets for $\Omega_{\mathcal{M}, k}(\cdot; \boldsymbol{\omega})$. We detail the construction of these sets for $\Omega_{\mathcal{M}, k}(\boldsymbol{\mu}; \boldsymbol{\omega})$ below, and the construction for $\Omega_{\mathcal{M}, k}(\mathbb{F}; \boldsymbol{\omega})$ follows similar lines. To this end, we consider the process

$\mathcal{V}_{\mathcal{M}, k}(\mathbf{x}) = \{\mathcal{V}_{km}(\mathbf{x})\}_{m \in \mathcal{M}}$, where $\mathcal{V}_{km}(\mathbf{x}) = (nh_k)^{1/2} \{\hat{\mu}_{km}(\hat{\mathbf{b}}_k^\top \mathbf{x}) - \mu_{km}(\bar{\mathbf{b}}_k^\top \mathbf{x})\}$. It follows

from similar arguments as given in Tian, Zucker and Wei (2005) and Cai et al. (2010) that for any $\|\mathcal{M}\|$ -variate vector $\mathbf{a}$, $\mathbf{a}^\top \mathcal{V}_{\mathcal{M}, k}(\mathbf{x})$ converges pointwise to a normal with some

variance function $\zeta_{\mathcal{M}, k}^2(\mathbf{x}; \mathbf{a})$ and properly standardized

$\sup_{\bar{\mathbf{b}}_k^\top \mathbf{x} \in [l_k + h_k, u_k - h_k]} |\mathbf{a}^\top \mathcal{V}_{\mathcal{M}, k}(\mathbf{x}) / \zeta_{\mathcal{M}, k}(\mathbf{x}, \mathbf{a})|$ converges to an extreme value distribution for

some l_k, u_k . Thus, $P\{|\mathcal{V}_{km}(\mathbf{x}) / \zeta_{km}(\mathbf{x})| \leq \mathcal{Z}_{1-\alpha/2}\} \approx 1 - \alpha$, where $\mathcal{Z}_\alpha = \Phi^{-1}(\alpha)$ and $\zeta_{km}(\mathbf{x})$ is the asymptotic variance function for $\mathcal{V}_{km}(\mathbf{x})$. Furthermore, there exists $\mathbf{d} = (d_m)_{1 \leq m \leq M}$

such that $P\{\sup_{\bar{\mathbf{b}}_k^\top \mathbf{x} \in [l_k + h_k, u_k - h_k]} |\mathcal{V}_{km}(\mathbf{x})| / \zeta_{km}(\mathbf{x}) \leq d_m, m \in \mathcal{M}\} \approx 1 - \alpha$. With $\mathbf{d}$

selected, we may identify simultaneous inner and outer confidence sets $\mathcal{N}_{\mathcal{M}, k}(\boldsymbol{\mu}; \boldsymbol{\omega}) = \{\mathbf{x} : \hat{\mu}_{km}(\hat{\mathbf{b}}_k^\top \mathbf{x}) - (nh_k)^{-1/2} d_m \mathcal{N} \zeta_{km}(\mathbf{x}) \leq \omega_m, m \in \mathcal{M}\}$ and $\mathcal{O}_{\mathcal{M}, k}(\boldsymbol{\mu}; \boldsymbol{\omega}) = \{\mathbf{x} : \hat{\mu}_{km}(\hat{\mathbf{b}}_k^\top \mathbf{x}) + (nh_k)^{-1/2} d_m \mathcal{O} \zeta_{km}(\mathbf{x}) \leq \omega_m, m \in \mathcal{M}\}$ having the property that $\mathcal{N}_{\mathcal{M}, k}(\boldsymbol{\mu}; \boldsymbol{\omega}) \subset \Omega_{\mathcal{M}, k}(\boldsymbol{\mu}; \boldsymbol{\omega})$ and $\mathcal{O}_{\mathcal{M}, k}(\boldsymbol{\mu}; \boldsymbol{\omega}) \supset \Omega_{\mathcal{M}, k}(\boldsymbol{\mu}; \boldsymbol{\omega})$ with probability approximately $1 - \alpha$.

To select $\mathbf{d}_{\mathcal{N}}$ and $\mathbf{d}_{\mathcal{O}}$ in practice, we may use resampling to approximate the distribution of $\mathcal{V}_{km}(\mathbf{x})$ along with the sup-statistics. For example, the distribution of $\mathcal{V}_{km}(\mathbf{x})$ may be approximated by that of

$$\mathcal{V}_{km}^*(\mathbf{x}) = (nh_k)^{1/2} \times \frac{\sum_{i=1}^n K_{h_k} \{ \Phi(\hat{\mathbf{b}}_k^\top \mathbf{x}) - \Phi(\hat{\mathbf{b}}_k^\top \mathbf{x}_i) \} \{ \mathbb{I}_{\tau_1, \tau_2}(y_{im}) - \hat{\mu}_{km}(\hat{\mathbf{b}}_k^\top \mathbf{x}) \} (\xi_1 - 1)}{\sum_{i=1}^n K_{h_k} \{ \Phi(\hat{\mathbf{b}}_k^\top \mathbf{x}) - \Phi(\hat{\mathbf{b}}_k^\top \mathbf{x}_i) \}}. \quad (10)$$

The variance function $\zeta_{km}(\mathbf{x})$ may be estimated as the empirical variance of $\mathcal{V}_{km}^*(\mathbf{x})$ conditional on the data, denoted by $\hat{\zeta}_{km}(\mathbf{x})$. Here, we ignore the variability of $\hat{\mathbf{b}}_k$ in estimating the variation of $\mathcal{V}_{km}(\mathbf{x})$ since $\mathcal{V}_{km}(\mathbf{x})$ has a slower convergence rate of $(nh_k)^{-1/2}(\log n)^2 + h_k^2$ while $\hat{\mathbf{b}}_k$ has a standard root- n convergence rate. The empirical distribution of $\mathcal{S}_{km}^* = \sup_{\hat{\mathbf{b}}_k^\top \mathbf{x} \in [l_k + h_k, u_k - h_k]} |\mathcal{V}_{km}^*(\mathbf{x}) / \hat{\zeta}_{km}(\mathbf{x})|$ can then be used to choose the cutoffs for the confidence sets. We may let $d_{m\mathcal{N}}$ be the $(1 - \alpha)^{th}$ percentile of $\max_{m \in \mathcal{M}} \mathcal{S}_{km}^*$ and $d_{m\mathcal{O}}$ be the $(1 - \alpha)^{th}$ percentile of $\min_{m \in \mathcal{M}} \mathcal{S}_{km}^*$.

4. Example and Simulation Studies

4.1. Genetic Study to Identify Risk Profiles for Autoimmune Disease

We apply sCCA to a study of shared autoimmunity with the goal of identifying genetic risk scores for predisposition to autoimmunity that crosses traditional phenotypic boundaries. The risk score is to be based on age, gender, and 67 single-nucleotide polymorphisms (SNPs) previously published as potential risk markers for four autoimmune diseases (ADs). The outcomes are four autoantibodies: anti-nuclear antibodies (ANA), anti-cyclic citrullinated protein (CCP) antibodies, anti-transglutaminase (TTG) antibodies, and anti-thyroid peroxidase antibodies (TPO). These four autoantibodies are respectively markers for 4 ADs: systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), celiac disease, and autoimmune thyroid disease. We seek to produce risk scores that are predictive of susceptibility to all or many ADs. These risk scores may hint at genetic etiology of autoimmune subtypes. Furthermore, based on these risk scores, patients at high risk of multiple ADs may potentially be identified prior to showing symptoms and selected for preventive care. We use the autoantibodies in place of the AD phenotypes because, while the co-occurrence of multiple ADs within individuals is not unheard of (Somers et al., 2006), it would be rare. Autoantibodies can be present in individuals predisposed to having the disease even in the absence of a disease phenotype and may indicate increased risk.

The study cohort includes 1265 individuals of European ancestry with RA identified through electronic medical records at Partners Healthcare (Liao et al., 2010). Due to limits of quantification, the true autoantibody measurements z_m are subject to truncation when the values are either very low or very high. We thus only analyze y_m , the truncated versions. All

predictors were standardized to have mean 0 and unit variance. The bootstrap with $\mathcal{J} = 3000$ was used to generate resampled estimates. Quantile-based CIs were used to estimate variability in $\hat{\theta}$. The `polr` function in the MASS R package was used to fit marginal models. The estimated transformed phenotypes are pictured in Figure 1. It is clear from Figure 1 that despite starting on very different scales, the transformed phenotypes are on similar scales.

All autoantibodies were estimated to be moderately positively correlated, with the exception that TPO was negatively correlated with ANA and CCP. The proportions of variance

explained by the first through fourth directions, $\hat{\mathbf{b}}_1^\top \mathbf{x}$, $\hat{\mathbf{b}}_2^\top \mathbf{x}$, $\hat{\mathbf{b}}_3^\top \mathbf{x}$, $\hat{\mathbf{b}}_4^\top \mathbf{x}$, were estimated to be 4.2% (bias-adjusted 95% CI 2.5–6.2%), 3.7% (2.5–5.0%), 1.9% (0.7–3.1%), and 1.1 (0.4–2.1%), respectively. The results for the first two sCCA directions, shown in Figure 2,

demonstrate that the first sCCA risk score $\hat{\mathbf{b}}_1^\top \mathbf{x}$ is driven primarily by a positive association with TPO and CCP and a negative association with TTG. The second direction is driven primarily by a negative association with CCP and a positive association with ANA and TPO.

Finally, to identify high-risk patients and examine the predictive utility of the risk scores, we plot the trimmed mean of each y_m smoothed against the first two risk profiles in Figure 3.

Values of y_m were trimmed at $\tau_{1m} = 0$ below and above at 10% higher than the highest observed value, $\tau_{2m} = 1.1 \times (\max_{\delta_{im}=1} y_{im})$. We undersmoothed to ensure the validity of the

bootstrap, choosing the smoothing parameter for y_m to be $h_k = \sqrt{\widehat{\text{var}}\{\Phi(\hat{\mathbf{b}}_k^\top \mathbf{x})\}} n^{-0.21}$.

Quantile-based confidence intervals were constructed. We recover associations similar to those that would be expected by examining the canonical directions. For example, we see in Figure 2 that the second sCCA direction was driven primarily by a strong association to ANA and TPO and a strong association in the opposite direction to CCP. Figure 3 confirms this – those with high values of the risk profile tend to have lower CCP levels and higher levels of other autoantibodies. Similarly, the first risk profile shows a strong positive association with CCP and TPO and a weaker association with ANA, though no negative association with TTG is apparent. This may be due to a degree of model misspecification in the TTG model, or it may be related to the low variation in TTG levels which could make the smoothed trimmed mean difficult to estimate. Considering the trends in Figure 3, the second direction can be seen to be picking up on CCP-negative autoimmunity – where patients with high values of the risk score tend to have higher values of most autoantibodies except CCP – and the first direction as CCP-positive autoimmunity – where patients with high values of the risk score tend to have higher values of all autoantibodies except TTG.

RA patients who are negative for CCP (values less than 20) are overrepresented among high values of the second risk score. About 22% (155/692) of individuals with scores less than 0 are negative for CCP, while about 42% (96/229) with scores above 1 and 50% (31/62) with scores above 2 are similarly negative.

To see what we gain by doing a joint analysis, we compare the prediction performance of

sCCA risk scores $\hat{\mathbf{b}}_k^\top \mathbf{x}_i$ to the prediction performance of the risk scores $\hat{\beta}_m^\top \mathbf{x}_i$ from the marginal models. Table 1 summarizes three measures of prediction performance, the rank correlation between the risk score and y_m , the AUC of the risk score for the dichotomized y_m at clinical cut-off ω_m , and the mean squared error of the smoothed trimmed mean $\hat{\mathbb{E}}_{\tau_1, \tau_2}$

$(\mathbf{x}, \hat{\mathbf{b}})$ or $\hat{\mathbb{E}}_{\tau_1, \tau_2}(\mathbf{x}, \hat{\mathbf{b}})$. The clinical cut-offs ω_m are: $\omega_1 = 20$ (weak positive ANA test), $\omega_2 = 60$ (strong positive CCP test), $\omega_3 = 100$ (positive TPO test), and $\omega_4 = 20$ (weak positive TTG test). There are a few things to take note of in table 1. First, note that the sCCA scores typically are related to two or three outcomes while the marginal risk scores are typically related only to one or maybe two. Second, the first sCCA score finds the “most predictable” outcomes (CCP and TPO) which have the highest marginal AUCs and rank correlations. Third, as noted above, the second risk score seems to pick out CCP-negative autoimmunity, finding a direction where those with high values of ANA and TPO tend to have low levels of CCP on average. Nothing like this is found when looking only at the marginal models. Fourth, it is interesting to note that the lowest MSE for each outcome is realized by one of the sCCA directions (for example, the lowest MSE for ANA is realized by the second sCCA direction).

Finally, we seek to identify subgroups of individuals who are at high risk for multiple ADs. Here, higher values of the autoantibodies correspond to greater risk, so we estimate the set

$\hat{\Omega}_{\{1,2,3\},1}(\hat{\mu}; \omega) = \{\mathbf{x}: \hat{\mu}_{1m}(\hat{\mathbf{b}}_1^\top \mathbf{x}) > \omega_m, m \in \{1, 2, 3\}\}$. In Figure 4 (a), the smaller shaded region corresponds to $\hat{\Omega}_{\{1,2,3\},1}(\hat{\mu}; \omega)$. All of the patients with values of $\mathbf{x}$ that lie in $\hat{\Omega}_{\{1,2,3\},1}(\hat{\mu}; \omega)$, have risk scores that lie in the shaded region. The larger shaded region corresponds to the outer confidence set. The inner confidence set for $\mathcal{M} = \{1, 2, 3\}$, however, is the empty set. For this reason, we also consider confidence sets for just ANA and CCP, depicted in Figure 4 (b). Here, the larger shaded region corresponds to the outer confidence set, the middle shaded region to $\hat{\Omega}_{\{1,2\},1}(\hat{\mu}; \omega)$, and the smallest shaded region to the inner confidence set.

4.2. Simulation Results

We performed simulations to investigate the finite-sample performance of sCCA. We generated data such that, for a given $\Sigma_{\varepsilon\varepsilon}$ and Σ_{xx}

$$z_m = h_m^*(\mathbf{x}^\top \beta_{m0} + \varepsilon_m), \varepsilon \sim \Phi_{\Sigma_{\varepsilon\varepsilon}}, \mathbf{x} \sim \Phi_{\Sigma_{xx}}, m=1, \dots, 4; p=1, \dots, 10. \quad (11)$$

We used $M = 4$ outcomes and $p = 10$ predictors. The transformations were chosen to be $h_1^*(x) = e^x$, $h_2^*(x) = x^3$, $h_3^*(x) = x$, and $h_4^*(x) = -\exp\{-e^x\}$. To demonstrate sCCA's ability to handle diverse $\mathbf{y}$, we thresholded y_2 and y_3 at 0, creating binary variables: $y_m = I\{z_m > 0\}$, $m = 2, 3$.

We considered sample sizes of $n = 250, 500$, and various covariance structures for $\Sigma_{\varepsilon\varepsilon}$, including independence ($\Sigma_{\varepsilon\varepsilon} = I$), exchangeable ($\Sigma_{\varepsilon\varepsilon} = 0.311I + 0.7J$), and unstructured. The form of the unstructured covariance matrix is given in web appendix B. The correlation matrix for $\mathbf{x}$, Σ_{xx} , was set to be the identity matrix. To demonstrate the finite-sample performance of the interval estimation, we performed 500 bootstrap samples for each simulation setting.

Overall, performance of sCCA estimates are quite good after bias correction, though estimates for later directions (here, directions 3 and 4) may tend to degrade somewhat, especially at the lower sample size. For ease of comparison across simulations, $\hat{\mathbf{a}}_k$ and $\hat{\mathbf{b}}_k$ were normalized to have norm 1.

We first examine the bias in our estimators. In Figure 5, we depict absolute bias. As expected, the biases tend to be higher for smaller sample sizes and for later directions. The bias in the estimation of the covariance between outcomes in $\hat{\Sigma}_{hh}$ is small regardless of correlation structure and is nearly nonexistent when $n = 500$. These results suggest that the proposed methods can provide reliable estimation for the leading directions at moderate sample sizes. Much of the bias depicted in Figure 5 is attenuated by using the bias-adjusted $\hat{\theta}_{BC}$. However, the estimates of later directions (i.e., the third and fourth canonical directions here) may still incur relatively large amounts of bias in practice, suggesting that inference may be less reliable for these directions and thus we focus our discussions primarily on the first few directions.

See web appendix B for further simulation results demonstrating the adequacy of bootstrap standard errors for sCCA estimates.

4.2.1. Comparison to VGLM—In this section, we compare the prediction performance of sCCA to that of reduced rank VGLM (rrVGLM) (Yee, 2015), as implemented in the VGAM package in R. The basic idea of rrVGLM is to find a low rank approximation of $\mathbb{B}$, the matrix of regression coefficients. While the VGLM framework is rich, no software has been implemented for mixed outcomes like the ones we are concerned with in this article. For this reason, we limit ourselves to the situation when $\mathbf{y}$ contains $M = 4$ binary variables.

We considered four different settings at $p = 16$ and three sample sizes, $n = 250, 500, 1000$. For each setting, we randomly generated Σ_{xx} , Σ_{ee} , and $\mathbb{B}$. As in the previous section, y_m is a thresholded version of $z_m = h_m^*(\mathbf{x}^\top \beta_{m0} + \varepsilon_m)$. See appendix B for full details about data generation. Across all configurations, sCCA is much better than rrVGLM with respect to out-of-sample (with 10,000 observations) prediction performance: (i) prediction of $\mathbf{y}$ measured by area under the receiver operating characteristic curve (AUC); (ii) prediction of the primary latent direction $\alpha_1^\top \mathbf{h}$ measured by rank correlation.

We summarize results for one setting here at $n = 250$, as a representative demonstration. See appendix B for full results on the other settings. Let $\hat{A}_m(\hat{\mathbf{b}})$ be the estimated AUC using $\hat{\mathbf{b}}^\top \mathbf{x}$ to predict y_m . Over 1000 simulations, the maximum AUC using the primary sCCA direction $\max_m \hat{A}_m(\hat{\mathbf{b}}_1)$ was on average 0.78 (empirical CI 0.72–0.82). Using the primary rrVGLM direction $\tilde{\mathbf{b}}_1$, $\max_m \hat{A}_m(\tilde{\mathbf{b}}_1)$ was 0.66 on average (empirical CI 0.59–0.73). On average, the primary sCCA direction produced a risk score that was predictive, defined as achieving $\hat{A}_m(\hat{\mathbf{b}}_1) > 0.65$, for 2.5 outcomes (100% of the time one outcome was well predicted, 98% two outcomes were, 46% three were, and 3% all four were), while the rrVGLM estimator was only predictive for 0.8 outcomes on average (61% one outcome was well predicted, 14% two were, 1% three were, and never were all four outcomes well predicted). Similarly, the absolute rank correlation between $\hat{\mathbf{b}}_1^\top \mathbf{x}$ and z_m was above 0.3 for two outcomes in 92%

of simulations, for three outcomes in 43%, and for all four outcomes in 1%. On the other hand, the absolute rank correlation between $\tilde{\mathbf{b}}_1^\top \mathbf{x}$ and z_m was above 0.3 for two outcomes in just 7% of simulations. The second and third sCCA directions were much more predictive than the best rrVGLM direction. The first sCCA direction $\hat{\mathbf{b}}_1^\top \mathbf{x}$ is also substantially more predictive than $\tilde{\mathbf{b}}_1^\top \mathbf{x}$ for the primary latent trait $\mathbf{a}_1^\top \mathbf{h}$. On average, $|\text{cor}(\hat{\mathbf{b}}_1^\top \mathbf{x}, \mathbf{a}_1^\top \mathbf{h})|$ was 0.74 (empirical CI 0.55–0.81), while $|\text{cor}(\tilde{\mathbf{b}}_1^\top \mathbf{x}, \mathbf{a}_1^\top \mathbf{h})|$ was 0.39 (empirical CI 0.15–0.60). Results for other simulation settings were quite similar.

5. Discussion

We have proposed sCCA, a method to synthesize diverse information in a set of phenotypes $\mathbf{y}$ for the purposes of joint prediction based on a set of predictors $\mathbf{x}$. That is, sCCA allows researchers to build risk profiles for many related phenotypes, even when those observed phenotypes are on incomparable scales, measured in completely different ways, or incompletely observed. Moreover, the canonical directions may offer insight into clustering and related subgroups in $\mathbf{x}$ and $\mathbf{y}$, and non-parametric techniques along with sCCA can be used to identify and make robust inference about the subgroups. We have shown that sCCA outperforms current state-of-the-art methods for handling multiple phenotypes such as VGLM.

When the number of predictors is moderately large, some regularization or variable selection may improve the performance of sCCA. To perform variable selection, the methods in Agniel, Liao and Cai (2016) may be used to induce sparsity in $\boldsymbol{\beta}$ when estimating the marginal models (4). The estimation strategy in Agniel, Liao and Cai (2016) induces sparsity at two levels. It can eliminate noninformative predictors altogether (by estimating $\hat{\beta}_{jm} = 0$, $m = 1, \dots, M$ simultaneously) or may set individual $\hat{\beta}_{jm}$ s to 0. This may improve interpretability and performance at lower sample sizes.

While variable selection is available for moderate dimensions, our sCCA methods as proposed above are not yet suitable to incorporate high-dimensional $\mathbf{x}$, since the performance of the model (4) and standard covariance matrix estimates degrade as the dimension of $\mathbf{x}$ becomes comparable to and exceeds the sample size n . However, there are many interesting clinical situations where considering high-dimensional $\mathbf{x}$ may be desirable, most notably in genomics where the number of potential markers can be quite large. In these cases, one may want to employ sparse estimators to improve sCCA, similar to what has been done in Gao, Ma and Zhou (2014) by first estimating the covariance matrices based on methods proposed in this article with L_2 regularization and then imposing regularization on $\mathbf{a}$ and $\mathbf{b}$ using these estimated covariance matrices. Similarly, there are many applications where high-dimensional $\mathbf{y}$ may be of interest. For example, $\mathbf{y}$ could conceivably contain the time until the first occurrence of each ICD-9 code in a patient's health record. In this case, penalization of the likelihood to obtain an estimate of Σ_{ee} in (8) may become necessary. These high-dimensional settings warrant further research.

Furthermore, when identifying high-risk patients, our method requires the researcher to select a set of outcomes $\mathcal{M} \subset \{1, \dots, M\}$ of primary interest for each risk score. In the event

of prior knowledge about $\mathcal{M}$, it might make sense to limit the study to just the outcomes in $\mathcal{M}$. However, in general we assume researchers will be operating without such prior knowledge, thus making the data-driven approach to developing traits even more appealing. One may rely on exploratory analyses to choose an $\mathcal{M}$ or even consider multiple choices of $\mathcal{M}$. Currently, for the data example, we chose $\mathcal{M}$ based on the subset of outcomes for which the risk score appears to be predictive given the clinical thresholds associated with the autoantibodies. One could potentially choose $\mathcal{M}$ as the set for which the inner (or outer) confidence set is not empty (the entire sample). This selection is a limitation of our approach, and a fully data-driven method for selecting $\mathcal{M}$ is a topic for future study.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by NIH grant P50 MH106933.

References

- Agniel D, Liao KP, Cai T. Estimation and testing for multiple regulation of multivariate mixed outcomes. *Biometrics*. 2016; 72:1194–1205. [PubMed: 26910481]
- Agresti, A., Kateri, M. *Categorical Data Analysis*. Heidelberg: Springer; Berlin: 2011. p. 206–208.
- Alarcón-Segovia D. Shared autoimmunity: A concept for which the time has come. *Autoimmunity*. 2005; 38:201–203. [PubMed: 16126507]
- Cai T, Tian L, Uno H, Solomon SD, Wei L. Calibrating parametric subject-specific risk estimation. *Biometrika*. 2010; 97:389–404. [PubMed: 23049123]
- Cai T, Tian L, Wei L. Semiparametric box–cox power transformation models for censored survival observations. *Biometrika*. 2005; 92:619–632.
- Doshi-Velez F, Ge Y, Kohane I. Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis. *Pediatrics*. 2014; 133:e54–e63. [PubMed: 24323995]
- Gao C, Ma Z, Zhou HH. Sparse cca: Adaptive estimation and computational barriers. *arXiv preprint arXiv:1409.8565*. 2014
- Glahn HR. Canonical correlation and its relationship to discriminant analysis and multiple regression. *Journal of the Atmospheric Sciences*. 1968; 25:23–31.
- Hardoon D, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*. 2004; 16:2639–2664. [PubMed: 15516276]
- Hotelling H. Relations between two sets of variates. *Biometrika*. 1936; 28:321–377.
- Jin J, Cui H. Asymptotic distributions in the projection pursuit based canonical correlation analysis. *Science China Mathematics*. 2010; 53:485–498.
- Jin Z, Ying Z, Wei L. A simple resampling method by perturbing the minimand. *Biometrika*. 2001; 88:381–390.
- Joe H. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*. 2005; 94:401–419.
- Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care & Research*. 2010; 62:1120–1127. [PubMed: 20235204]
- Lin H, Zhou L, Elashoff R, Li Y. Semiparametric latent variable transformation models for multiple mixed outcomes. *Statistica Sinica*. 2013; 24:833–854.
- Michailidis G, de Leeuw J. The gifi system of descriptive multivariate analysis. *Statistical Science*. 1998; 13:307–336.

- Murphy SA, Van der Vaart AW. On profile likelihood. *Journal of the American Statistical Association*. 2000; 95:449–465.
- O'Brien WA, Hartigan PM, Martin D, Esinhart J, Hill A, Benoit S, et al. Changes in plasma HIV-1 RNA and CD4+ lymphocyte counts and the risk of progression to AIDS. *New England Journal of Medicine*. 1996; 334:426–431. [PubMed: 8552144]
- Ogasawara H. Asymptotic expansions of the distributions of estimators in canonical correlation analysis under nonnormality. *Journal of Multivariate Analysis*. 2007; 98:1726–1750.
- Othuis M, Li Y. A gaussian copula model for multivariate survival data. *Statistics in Biosciences*. 2010; 2:154–179. [PubMed: 22162742]
- Snavely AC, Harrington DP, Li Y. A latent variable transformation model approach for exploring dysphagia. *Statistics in Medicine*. 2014; 33:4337–4352. [PubMed: 24974798]
- Somers EC, et al. Autoimmune diseases co-occurring within individuals and within families: A systematic review. *Epidemiology*. 2006; 17:202–217. [PubMed: 16477262]
- Song PXX, Li M, Yuan Y. Joint regression analysis of correlated data using gaussian copulas. *Biometrics*. 2009; 65:60–68. [PubMed: 18510653]
- Tian L, Zhao L, Wei L. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*. 2014; 15:222–233. [PubMed: 24292992]
- Tian L, Zucker D, Wei L. On the cox model with time-varying regression coefficients. *Journal of the American Statistical Association*. 2005; 100:172–183.
- Venables, WN., Ripley, BD. *Modern Applied Statistics with S*. 4th. Springer; New York: 2002.
- Wand MP, Marron JS, Ruppert D. Transformations in density estimation. *Journal of the American Statistical Association*. 1991; 86:343–353.
- Xia Y. A semiparametric approach to canonical analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2008; 70:519–543.
- Yee, TW. *Vector Generalized Linear and Additive Models: With an Implementation in R*. New York: Springer-Verlag; 2015.
- Zeng D, Lin D. A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Statistica Sinica*. 2010; 20:871. [PubMed: 20577580]
- Zhou L, Lin H, Song X, Li Y. Selection of latent variables for multiple mixed-outcome models. *Scandinavian Journal of Statistics*. 2014; 41:1064–1082.
- Zhu LX, Zhu LP, Li X. Transformed partial least squares for multivariate data. *Statistica Sinica*. 2007; 17:1657.

Figure 1.
Estimated transformation function $\hat{h}(\cdot)$ for the four autoantibodies.

Figure 2.

First two canonical directions for $\mathbf{x}$ and $\mathbf{h}$ in the autoantibody study. Dots indicate point estimates. Line ranges represent bootstrap estimates of 95% CIs.

Figure 3. Nonparametric smoothing of trimmed mean against risk score. Quantile-based bootstrap confidence intervals are given by the broken lines.

Figure 4. Identification of high-risk patients. The solid lines represent the estimated trimmed means $\hat{\mu}_{1m}(\hat{\mathbf{b}}_1^\top \mathbf{x}; 0, \tau_{2m})$, the lower dashed lines represent $\hat{\mu}_{1m}(\hat{\mathbf{b}}_1^\top \mathbf{x}; 0, \tau_{2m}) - (nh_1)^{-1/2} d_{m\mathcal{N}} \hat{\zeta}_{1m}(\mathbf{x})$, and the upper broken lines represent $\hat{\mu}_{1m}(\hat{\mathbf{b}}_1^\top \mathbf{x}; 0, \tau_{2m}) + (nh_1)^{-1/2} d_{m\mathcal{O}} \hat{\zeta}_{1m}(\mathbf{x})$. The horizontal dotted lines represent the cutoffs ω_m , $m = 1, 2, 3$. The largest shaded region corresponds to the outer 95% confidence set $\mathcal{O}_{\mathcal{M},1}$ and the smaller shaded region to the point estimate of the high-risk subgroup $\hat{\mathcal{Q}}_{\mathcal{M},1}$. In (b), the smallest shaded region corresponds to the inner confidence set $\mathcal{N}_{\mathcal{M},1}$ (in (a) the inner confidence set is the empty set).

Figure 5.

Absolute bias of sCCA estimators, represented by the height of bars. Row panel denotes simulation setting, that is, correlation structure of $\mathbf{e}$. Column panels represent small and large sample sizes as well as unadjusted and bias-adjusted estimators. The x-axis denotes canonical direction (k) for $\hat{\alpha}_k$, $\hat{\beta}_k$, and $\hat{\gamma}_k$, while for $\hat{\Sigma}_{hh}$ the x-axis denotes the row in the matrix, and individual bars represent the bias at entries in the vector.

Table 1
Results of comparing risk scores based on sCCA ($\hat{\mathbf{b}}_k$) and risk scores based on marginal models ($\hat{\boldsymbol{\beta}}_k$).

Coefficients	Rank correlation				AUC				MSE			
	ANA	CCP	TPO	TTG	ANA	CCP	TPO	TTG	ANA	CCP	TPO	TTG
$\hat{\boldsymbol{\beta}}_1$	0.24	0.02	0.14	0.06	0.58	0.51	0.58	0.52	1461	40,326	74,090	70
$\hat{\boldsymbol{\beta}}_2$	0.01	0.35	-0.01	0.04	0.51	0.64	0.51	0.51	1557	41,447	74,262	97
$\hat{\boldsymbol{\beta}}_3$	0.11	0.01	0.27	0.02	0.54	0.50	0.66	0.51	1486	40,599	74,885	72
$\hat{\boldsymbol{\beta}}_4$	0.08	0.03	0.04	0.23	0.53	0.51	0.52	0.59	1495	39,891	73,944	70
$\hat{\mathbf{b}}_1$	0.07	0.24	0.17	-0.01	0.52	0.59	0.59	0.51	1487	38,856	72,365	73
$\hat{\mathbf{b}}_2$	0.15	-0.23	0.19	0.00	0.55	0.59	0.61	0.50	1453	41,511	74,074	70
$\hat{\mathbf{b}}_3$	0.11	0.07	-0.01	0.22	0.54	0.53	0.51	0.59	1457	39,506	73,027	68
$\hat{\mathbf{b}}_4$	-0.10	-0.07	0.09	0.10	0.54	0.53	0.55	0.54	1513	40,673	74,858	72