*The Annals of Applied Statistics*
2019, Vol. 13, No. 1, 321–339
https://doi.org/10.1214/18-AOAS1187
© Institute of Mathematical Statistics, 2019

# JOINT MEAN AND COVARIANCE MODELING OF MULTIPLE HEALTH OUTCOME MEASURES

BY XIAOYUE NIU[1] AND PETER D. HOFF[2]

*Pennsylvania State University and Duke University*

Health exams determine a patient's health status by comparing the patient's measurement with a population reference range, a 95% interval derived from a homogeneous reference population. Similarly, most of the established relation among health problems are assumed to hold for the entire population. We use data from the 2009–2010 National Health and Nutrition Examination Survey (NHANES) on four major health problems in the U.S. and apply a joint mean and covariance model to study how the reference ranges and associations of those health outcomes could vary among subpopulations. We discuss guidelines for model selection and evaluation, using standard criteria such as AIC in conjunction with posterior predictive checks. The results from the proposed model can help identify subpopulations in which more data need to be collected to refine the reference range and to study the specific associations among those health problems.

**1. Introduction.** Health exams determine a patient's health status by comparing the patient's measurement with a population reference range. For example, in measuring blood sugar levels, the normal range of a fasting glucose level is 70 to 100 mg/dl. People with values lower than 70 mg/dl are considered to have hypoglycemia (low blood sugar), and people with values higher than 100 mg/dl are considered pre-diabetic (100–125 mg/dl) or diabetic (>125 mg/dl). The reference range is usually a 95% interval derived from a reference population. Current guidelines suggest that if the reference population is heterogeneous, we should partition it and provide a separate reference range for each subpopulation [CLSI (2008)]. Mattix et al. (2002) argue that using a single cutpoint in diagnosing kidney disease for both genders and various race groups biases the prevalence of the disease for some subpopulations and thus underestimates their risks. The most widely used partition guideline is that if the ratio of the two subpopulation standard deviations is greater than 1.5, we should collect large enough samples in those groups and provide separate reference ranges, regardless of whether the mean difference is significant or not [Harris and Boyd (1990)].

Similarly, certain health problems are associated with others. For example, obesity is a risk factor of diabetes. Most established relations among health problems

Received July 2015; revised January 2018.

[1] Supported by NIH Grant R01 AI36664-01.

[2] Supported by NSF Grant DMS-15-05136.

*Key words and phrases.* Heterogeneous population, reference range, covariance regression, NHANES.

are assumed to hold for the entire population. However, those relations could vary among subpopulations. For example, Foulds, Bredin and Warburton (2012) find that the relation between obesity and diabetes varies with ethnicity.

We use data from the 2009–2010 National Health and Nutrition Examination Survey (NHANES) to look at some major health problems in the U.S. [CDC/NCHS (2010a), http://www.cdc.gov/nchs/nhanes/search/nhanes09_10.aspx]. NHANES is designed to assess the health and nutritional status of adults and children in the United States. It collects participants' demographic, socioeconomic, dietary, activity, and behavioral information through interviews in their homes. It also performs physical measurements and blood and urine tests in mobile examination centers. The National Center for Health Statistics (NCHS), part of the Centers for Disease Control and Prevention (CDC), conducts the survey mainly to determine the prevalence of major diseases and risk factors in the U.S. population.

We focus on four health problems that are believed to be associated: chronic kidney disease (CKD), obesity, hypertension, and diabetes. The severity and progression of each health problem can be assessed by a quantitative measurement. In chronic kidney disease, defined as abnormalities of kidney structure or function, the kidneys are damaged and cannot filter blood as needed. Kidney damage and disease progression can be assessed by the urine albumin/creatinine ratio (ACR). ACR below 30 mg/g is considered normal and above 30 mg/g is considered to indicate microalbuminuria, a marker for CKD and kidney damage [KDIGO (2013)]. Obesity is quantified by body mass index (BMI), defined as weight in kilograms divided by the square of height in meters. For adults of 20 years and older, a BMI below 18.5 is considered underweight, 18.5–24.9 is normal, 25–29.9 is overweight, and over 30 is obese. Hypertension (high blood pressure) is diagnosed by measuring both systolic blood pressure (SBP) and diastolic blood pressure (DBP). A normal blood pressure corresponds to SBP of 80–120 mmHg AND DBP of 60–80 mmHg. If SBP $>$ 120 or DBP $>$ 80, a patient is considered to have elevated blood pressure. If SBP $>$ 120 and DBP $>$ 80, a patient is considered to have hypertension. If SBP $<$ 80 or DBP $<$ 60, a patient is considered to have hypotension (low blood pressure). SBP and DBP are usually correlated, so we take DBP to represent blood pressure (BP). Finally, a common measurement for diabetes is the fasting glucose level (GLU), discussed earlier.

All of these reference ranges are derived by assuming that the measurement comes from a homogeneous population, summarized by its mean and variance. Based on the CLSI guidelines and some previous findings [NIDDK (2013), Fraser et al. (2012)], we use gender, age, race/ethnicity, and education level to to define subpopulations. Refining the reference ranges and associations of the four health problems requires estimation of the means and covariances of ACR, BMI, BP, and GLU in the subpopulations. To estimate how the mean and covariance structure vary among subpopulations, we jointly model them as functions of the demographic variables.

Joint regression models for means and covariances have been developed mainly in the context of longitudinal and repeated-measures studies. Liang and Zeger (1986) and Zeger and Liang (1986) use generalized estimating equations (GEE) to simultaneously estimate the parameters in the mean and covariance of a longitudinal response vector, which improves the efficiency of the mean estimate substantially. When the heteroscedasticity is temporal, multivariate autoregressive conditionally heteroscedastic (ARCH) models are well studied in the econometric literature [Engle and Kroner (1995), Fong, Li and An (2006)]. The approach proposed by Pourahmadi (1999) uses the Cholesky decomposition to parameterize the class of positive-definite covariance matrices by expressing the unconstrained parameters through generalized linear models. However, this model is not invariant to reorderings of the response, and thus might not be appropriate for studies without longitudinal or spatial structure. Chiu, Leonard and Tsui (1996) model the logarithm of the covariance matrix as linear functions of the explanatory variables, although the parameters are somewhat difficult to interpret. Pourahmadi (2011) gives a comprehensive literature review of covariance estimation models.

Hoff and Niu (2012) propose a covariance regression model that directly models the covariance matrix as a function of explanatory variables. In this natural extension of the mean regression model, the parameters have interpretations similar to those in a mean regression. However, Hoff and Niu (2012) focus on model development and geometric interpretations. They discuss an example of a single continuous predictor. We extend the covariance regression model of Hoff and Niu (2012) to accommodate multiple categorical predictor variables. We also discuss practical issues with real data, including model selection and how to present and interpret the results.

In the next section, we explore some basic features of the NHANES data. In Section 3 we introduce the proposed method for joint modeling and outline the process of model selection. We describe the details of model selection and present our main findings in Section 4. Discussion follows in Section 5.

**2. The NHANES data.** The 2009–2010 NHANES had 10,537 participants, but only 3386 had a fasting glucose blood test. Among them, only 2613 (77%) had data for all four of ACR, BMI, BP (DBP), and GLU. Because GLU is an important measurement when dealing with kidney diseases, we use in this analysis only those who have complete data; the sample size reduction is due mainly to the small number of participants who took the blood tests (GLU). Further discussion of the sample size and missing data is in Section 5.

The demographic variables are categorical: gender (male, female), age (20–39, 40–59, 60–79, 80+), race/ethnicity (in order of decreasing sample size: non-Hispanic white, Mexican American, non-Hispanic black, other Hispanic, and other), education (less than 9th grade, 9th to 11th grade, high school, associate degree or some college, and college degree and higher). For the 16 marginal groups defined by one category of one predictor, such as male, the sample sizes all exceed

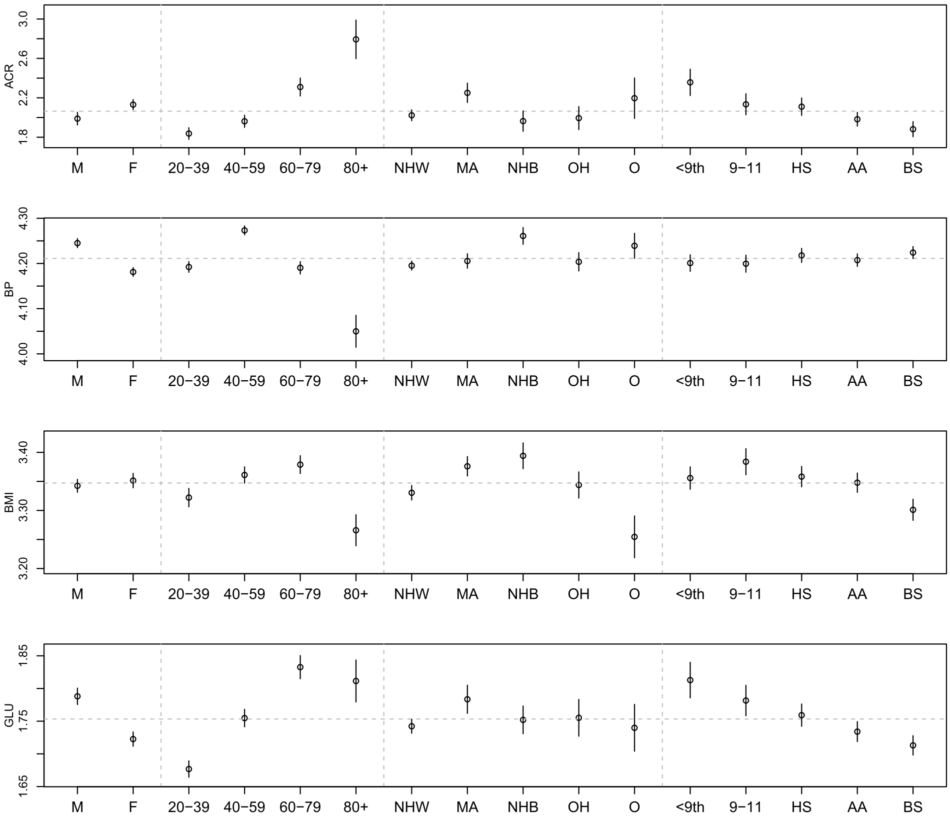

FIG. 1. *Sample means with* 95% *confidence intervals of albumin creatinine ratio* (*ACR*), *diastolic blood pressure* (*BP*), *body mass index* (*BMI*), *and glucose* (*GLU*), *on the natural log scale*, *by gender*, *age*, *race/ethnicity* (*NHW*: *non-Hispanic white*, *MA*: *Mexican American*, *NHB*: *non-Hispanic black*, *OH*: *other Hispanic*, *O*: *other*), *and education* (*<*9*th*: *less than* 9*th grade*, 9–11: 9*th to* 11*th grade*, *HS*: *high school*, *AA*: *associate degree or some college*, *BS*: *college degree and higher*) *categories*. *The horizontal dotted line indicates the pooled sample mean*. *The figure is based on the subset of* 2613 *individuals from NHANES* 2009–2010 *who have complete observations on these four variables*.

100. The sample sizes among the 93 two-way cells range from 4 to 660 (median 132). Of the 200 four-way cells, 28 are empty, and the median sample size among the other 172 four-way cells is 7.5 (range 1–74; quartiles 2 and 17). A detailed sample size tabulation is in Supplementary Material A [Niu and Hoff (2019)].

Because most of them are skewed, we analyze these variables on the natural log scale. In Figure 1, the sample means of the health measurements vary greatly among demographic groups. We also calculate the sample covariance matrices within the demographic groups, along with Bayesian posterior intervals using a noninformative Wishart prior. Figure 2 and Figure 3 show that the variances and correlations also vary among subpopulations. For example, the variance of ACR is about 1.14 overall, but it can be as low as 0.79 in the 60–79 age group and as

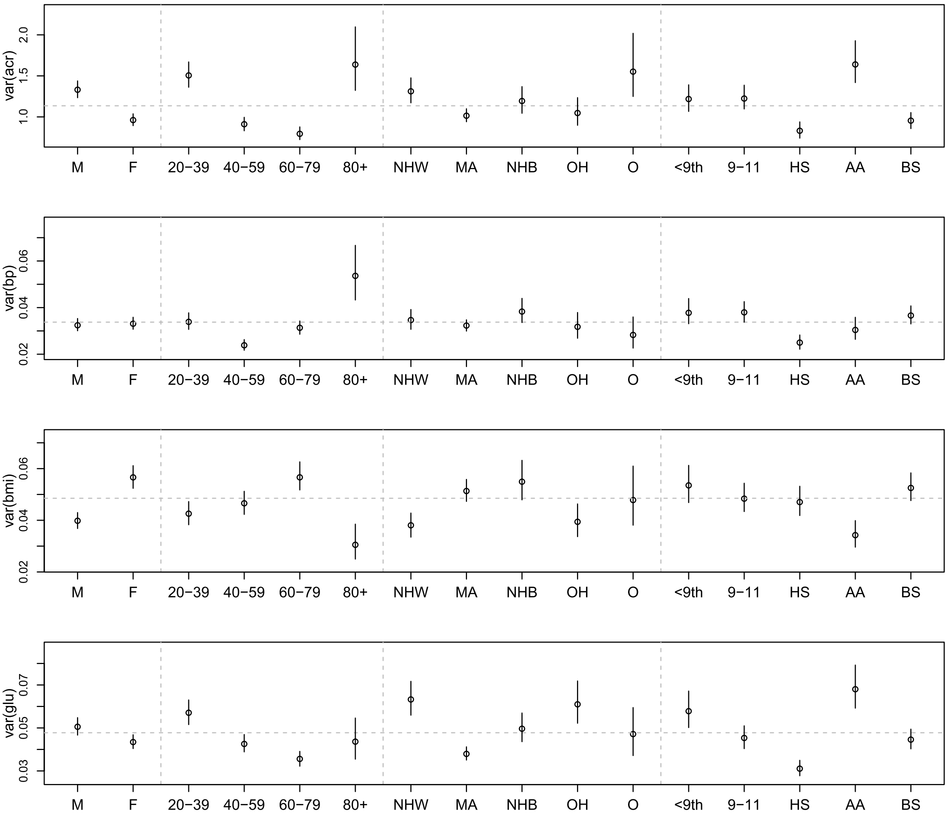

FIG. 2. *Sample variances with* 95% *Bayesian posterior intervals of albumin creatinine ratio* (*ACR*), *diastolic blood pressure* (*BP*), *body mass index* (*BMI*), *and glucose* (*GLU*), *on the natural log scale*, *by gender*, *age*, *race/ethnicity*, *and education categories*. *The horizontal dotted line indicates the pooled sample variance*. *The figure is based on the subset of* 2613 *individuals from NHANES* 2009–2010 *who have complete observations on these four variables*.

high as 1.64 in the 80+ age group. Similarly, the correlation between BP and GLU is around 0.06 overall, but it can be as low as −0.13 in the 60–79 age group and as high as 0.2 in the 80+ age group. We have also considered multiplicity of the intervals. Supplementary Material B [Niu and Hoff (2019)] includes Bonferroni-corrected simultaneous 95% intervals. The patterns are very similar to Figures 1 to 3. The exploratory findings suggest the need for a statistical analysis that allows both the mean and covariance matrix of these health outcomes to vary among demographic groups.

## 3. Statistical models.

3.1. *The covariance regression model.* Our goal is to describe and estimate heterogeneity of means and covariances for the cross-classified groups defined by

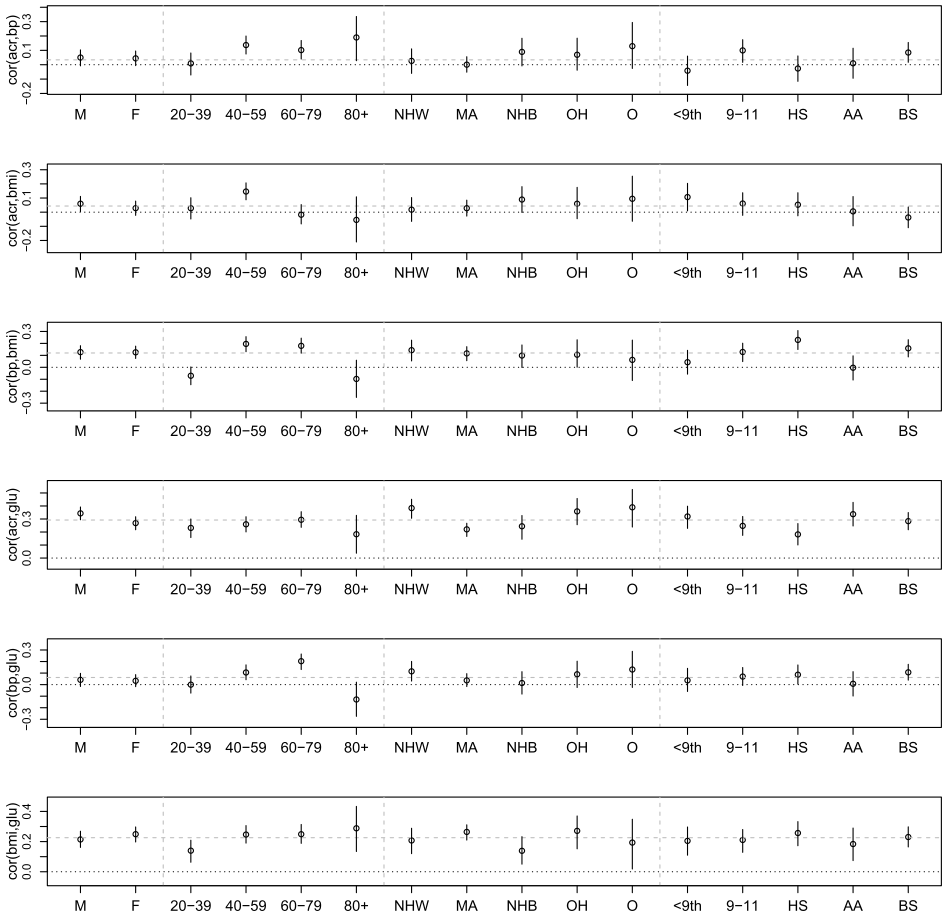

FIG. 3. *Sample pairwise correlations, with* 95% *Bayesian posterior intervals, for each pair of albumin creatinine ratio* (*ACR*), *diastolic blood pressure* (*BP*), *body mass index* (*BMI*), *and glucose* (*GLU*), *on the natural log scale, by gender, age, race/ethnicity, and education categories. The horizontal dotted line is at* 0. *The horizontal dashed line indicates the all-sample correlation. The figure is based on the subset of* 2613 *individuals from NHANES* 2009–2010 *who have complete observations on these four variables.*

gender, age, race/ethnicity, and education. Specifically, let $\mathbf{y}$ be the 4-dimensional vector of the logarithms of ACR, BP, BMI, and GLU of an individual, and let $\mathbf{x}$ be a covariate vector describing the individual's gender, age, race/ethnicity, and education level. We would like to estimate $\mathrm{E}[\mathbf{y}|\mathbf{x}] = \boldsymbol{\mu}_{\mathbf{x}}$ and $\mathrm{Cov}[\mathbf{y}|\mathbf{x}] = \boldsymbol{\Sigma}_{\mathbf{x}}$ simultaneously.

The small number of observations for each combination of categories makes it impractical to estimate a separate covariance matrix $\boldsymbol{\Sigma}_{\mathbf{x}}$ for each group, based on data from only that group. On the other hand, a common covariance matrix for all groups would misrepresent the relations among the response variables and result in loss of efficiency for the mean parameters (see McCullagh and Nelder (1989), Chapter 9 and Chapter 10). More-flexible models share information across covariance matrices. Boik (2002, 2003) assumes common principal components of the covariance matrix. Hoff (2009) proposes shrinking the covariance matrix toward a common eigenvector structure with varying degrees of shrinkage among principal components. Cripps, Carter and Kohn (2005) show efficiency improvement for the mean regression parameters with a carefully selected covariance model. Gaskins and Daniels (2013) give a more-comprehensive review of pooling methods.

Besides avoiding the loss of efficiency from assuming a common covariance structure, we are also interested in how $\mathbf{\Sigma}_{\mathbf{x}}$ varies with $\mathbf{x}$. An alternative way of pooling across groups addresses this question with a covariance regression model that parsimoniously describes heteroscedasticity among groups [Hoff and Niu (2012)]. The proposed model parametrizes the mean and covariance of a multivariate response vector as parsimonious functions of explanatory variables. This approach allows joint modeling of the mean and covariance structure of the population being studied.

To introduce the model, we adopt the notation and definitions in Hoff and Niu (2012). Let $\mathbf{y} \in \mathbb{R}^p$ be a random multivariate response vector and $\mathbf{x}_1 \in \mathbb{R}^{q_1}$ and $\mathbf{x}_2 \in \mathbb{R}^{q_2}$ be vectors of explanatory variables. The variables in $\mathbf{x}_1$ and $\mathbf{x}_2$ can overlap or even be the same. Denote the mean of $\mathbf{y}|\mathbf{x}$ as $\boldsymbol{\mu}_{\mathbf{x}} = \mathrm{E}[\mathbf{y}|\mathbf{x}]$ and the $p \times p$ covariance matrix of $\mathbf{y}|\mathbf{x}$ as $\mathbf{\Sigma}_{\mathbf{x}} = \mathrm{Cov}[\mathbf{y}|\mathbf{x}]$. The covariance regression model has the form

$$\boldsymbol{\mu}_{\mathbf{x}_1} = \mathbf{B}_1\mathbf{x}_1, \tag{3.1}$$

$$\mathbf{\Sigma}_{\mathbf{x}_2} = \mathbf{A} + \mathbf{B}_2\mathbf{x}_2\mathbf{x}_2^T\mathbf{B}_2^T, \tag{3.2}$$

where $\mathbf{B}_1$ is a $p \times q_1$ matrix, $\mathbf{A}$ is a $p \times p$ positive-definite matrix, and $\mathbf{B}_2$ is a $p \times q_2$ matrix. The resulting covariance function in equation (3.2) is positive-definite for all $\mathbf{x}_2$, and it expresses the covariance as a constant covariance matrix $\mathbf{A}$ plus a rank-1, positive-semi-definite matrix that varies with $\mathbf{x}_2$. Hoff and Niu (2012) consider the case where $\mathbf{x}_2$ is a single continuous variable, which makes the variance a quadratic function of the predictor. We apply such a model when there are multiple categorical predictors. For example, if sex is our only predictor, we let $\mathbf{x}_2 = (1, 1)^T$ for males and $\mathbf{x}_2 = (1, 0)^T$ for females, where the first "1" in each vector corresponds to the intercept.

The covariance regression model can also be interpreted as a special random-effects model. Assume the observed data $\mathbf{y}_1, \dots, \mathbf{y}_n$ are generated by the following model:

$$\begin{aligned} &\mathbf{y}_i = \boldsymbol{\mu}_{\mathbf{x}_{1i}} + \gamma_i \times \mathbf{B}_2\mathbf{x}_{2i} + \boldsymbol{\varepsilon}_i, \\ &\mathrm{E}[\boldsymbol{\varepsilon}_i] = \mathbf{0}, \qquad \mathrm{Cov}[\boldsymbol{\varepsilon}_i] = \mathbf{A}, \\ &\mathrm{E}[\gamma_i] = 0, \qquad \mathrm{Var}[\gamma_i] = 1, \qquad \mathrm{E}[\gamma_i \times \boldsymbol{\varepsilon}_i] = \mathbf{0}. \end{aligned} \tag{3.3}$$

We can interpret $\gamma_i$ as describing additional individual-level variability beyond the random error $\boldsymbol{\varepsilon}_i$. The row vectors $\{\boldsymbol{b}_{21}, \dots, \boldsymbol{b}_{2p}\}$ of the coefficient matrix $\mathbf{B}_2$ describe how this additional variability is manifested in the $p$ response variables.

Model (3.2) restricts the difference between $\boldsymbol{\Sigma}_{\mathbf{x}}$ and the constant matrix $\mathbf{A}$ to be a rank-1 matrix. This rank-1 model essentially requires that the residuals of the $p$ responses are along the same direction. This restriction can be relaxed by allowing the difference from the constant covariance to have higher rank. For example, a rank-2 covariance regression model has the following form:

$$\mathbf{y}_i = \boldsymbol{\mu}_{\mathbf{x}_{1i}} + \gamma_i \times \mathbf{B}_2 \mathbf{x}_{2i} + \psi_i \times \mathbf{B}_3 \mathbf{x}_{2i} + \boldsymbol{\varepsilon}_i, \tag{3.4}$$

where $\gamma_i$ and $\psi_i$ are mean-zero variance-one random variables, uncorrelated with each other and with $\boldsymbol{\varepsilon}_i$. $\mathbf{B}_2$ in equation (3.4) has a different estimate and interpretation than the $\mathbf{B}_2$ in equation (3.3) because of the additional term in equation (3.4). We keep the same notation for simplicity. Under this model, the covariance matrix of $\mathbf{y}_i$ is given by

$$\boldsymbol{\Sigma}_{\mathbf{x}_2} = \mathbf{A} + \mathbf{B}_2 \mathbf{x}_2 \mathbf{x}_2^T \mathbf{B}_2^T + \mathbf{B}_3 \mathbf{x}_2 \mathbf{x}_2^T \mathbf{B}_3^T. \tag{3.5}$$

Model (3.5) allows the deviation of $\boldsymbol{\Sigma}_{\mathbf{x}_2}$ from the constant matrix $\mathbf{A}$ to have rank 2. We can interpret the second random effect $\psi$ in equation (3.4) as allowing an additional, independent source of heteroscedasticity for the $p$ response variables. Further flexibility can be gained with additional random effects, allowing the difference between $\boldsymbol{\Sigma}_{\mathbf{x}}$ and the constant matrix $\mathbf{A}$ to be of any desired rank up to $p$.

Assuming normality of the error terms, the rank-1 model can be expressed as follows:

$$\begin{aligned}
\gamma_1, \dots, \gamma_n &\overset{\text{i.i.d.}}{\sim} \text{normal}(0, 1), \\
\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_n &\overset{\text{i.i.d.}}{\sim} \text{multivariate normal}(\mathbf{0}, \mathbf{A}), \\
\mathbf{y}_i &= \boldsymbol{\mu}_{\mathbf{x}_{1i}} + \gamma_i \times \mathbf{B}_2 \mathbf{x}_{2i} + \boldsymbol{\varepsilon}_i, \\
\boldsymbol{\mu}_{\mathbf{x}_{1i}} &= \mathbf{B}_1 \mathbf{x}_{1i}.
\end{aligned} \tag{3.6}$$

Parameters of this normal covariance regression model can be estimated by maximum likelihood via the EM algorithm or by Bayesian estimation via Markov chain Monte Carlo (MCMC). We focus on Bayesian estimation mainly for three reasons: 1. the convergence of EM can be very slow due to the identifiability issue; 2. it is easier to obtain the intervals for those identifiable parameters in the Bayesian setting; and 3. it is easier to perform model selection and diagnoses (such as posterior predictive checks discussed below). Calculations are facilitated by using a semi-conjugate prior distribution for $\mathbf{A}$, $\mathbf{B}_1$ and $\mathbf{B}_2$, in which $p(\mathbf{A})$ is an inverse-Wishart $(\mathbf{A}_0^{-1}, \nu_0)$ distribution and $\mathbf{C} = (\mathbf{B}_1, \mathbf{B}_2)$ has a matrix normal distribution.

Hoff and Niu (2012) introduce this covariance regression model in the context of continuous predictors, and give an example with one continuous predictor. We extend that model to jointly estimate the mean and covariance structure of a large number of groups defined by several categorical variables. We fit the model in equation (3.1) and (3.2) and its higher-rank version to the NHANES data, allowing different sets of predictors for the mean and covariance matrix. Including multiple categorical predictors requires that we address practical issues such as variable selection and evaluation, which are not discussed in Hoff and Niu (2012). In the next section, we discuss the outline of model selection for this covariance regression model with multiple categorical factors.

3.2. *Model selection and evaluation.* Similar to any regression model, the covariance regression model requires a procedure for variable selection. The process has three components: mean variable selection, covariance variable selection, and covariance rank selection. As noted in Hoff and Niu (2012), because of the nonidentifiability of some of the parameters in the higher-rank model, methods such as AIC or BIC are not directly applicable when comparing models with different ranks. For the NHANES data, we include 4 predictors with at most 2-way interactions (6 interaction terms) in both the mean and covariance models. The maximum possible rank for a 4-dimensional response is 4. Simultaneous selection of the appropriate interaction terms for the mean model and the covariance model and the selection of rank would require $2^6 \times 2^6 \times 4$ evaluations of the model, which is computationally impractical.

As an alternative, we propose a "forward search procedure" that tries to find the most parsimonious model without obvious lack of fit. We outline the procedure in this section and elaborate the details with data in Section 4. First, we simplify the situation by separating the tasks of mean and covariance model selection, based on the fact that under multivariate normality, the maximum likelihood estimator of the mean parameters is consistent under mis-specification of the covariance structure [Cox and Reid (1987)]. For selecting the mean model, we assume a homogeneous covariance model and use a standard variable-selection criterion such as AIC or BIC. Next we fix that mean model and fit the simplest covariance model, a rank-1 model with only main effects of the four predictors. Then we assess goodness of fit. If the simplest model has only moderate lack of fit, we add one interaction term at a time until we find a model that is acceptable. If the simplest model displays serious lack of fit, we increase the rank by 1 and implement forward selection for the rank-2 model until we find an acceptable set of predictors. If necessary, we can continue the selection to rank 3 or higher.

Model fit is evaluated with posterior predictive distributions [Guttman (1967) and Rubin (1984)]. To assess the model, we need to construct a meaningful statistic to represent lack of fit. We would like to make sure that the model we select generates predictive datasets $\tilde{\mathbf{Y}}$ that resemble the observed dataset $\mathbf{Y}$ (the observed response matrix) in terms of features that are of interest. The key idea is that the population is not homogeneous, and the covariance matrices differ among the subpopulations defined by the variables that make up $\mathbf{x}_2$. Therefore we construct a diagnostic statistic that describes the heteroscedasticity across subpopulations defined by pairwise combinations of the factors, such as all the white females, or people 30–49 years old with a high school degree. We use 2 variables instead of 4 because the sample size is not large enough for estimation of all groups obtained by cross-classifying the 4 variables. We define the posterior check statistic

$$t_{hk}(\mathbf{Y}) = \sum_{x_h,x_k} [\mathrm{tr}(\mathbf{S}_0^{-1}\mathbf{S}_{x_h,x_k}) - \log|\mathbf{S}_0^{-1}\mathbf{S}_{x_h,x_k}|],$$

where $\mathbf{S}_0$ is the all-sample covariance matrix, $h$ and $k$ denote the factors (e.g., gender and age), $x_h$ and $x_k$ represent the levels of factors $h$ and $k$ (e.g., male and 20–39 years old), and $\mathbf{S}_{x_h,x_k}$ is the sample covariance matrix of the subpopulation defined by $x_h$ and $x_k$ (e.g., males 20–39 years old). The statistic $t_{hk}(\mathbf{Y})$ is the sum of the Wishart kernels of the sample covariance matrices $\mathbf{S}_{x_h,x_k}$ for all possible values of $x_h$ and $x_k$, with $\mathbf{S}_0$ as the center. If the population is homogeneous, $\mathbf{S}_0$ should be a good estimate of the $\mathbf{S}_{x_h,x_k}$. The statistic $t_{hk}(\mathbf{Y})$ describes the discrepancy between $\mathbf{S}_0$ and the $\mathbf{S}_{x_h,x_k}$ and represents the heterogeneity of the subpopulation covariance matrices. For each model, we compute the posterior predictive distribution of $t_{hk}$ for all pairs of factors. We then compare the observed value with the posterior predictive distribution of $t_{hk}$. If the observed statistic lies in the tail of the posterior predictive distribution, it indicates lack of fit in that pair of factors.

## 4. Analysis of the NHANES data.

In this section, we first describe the detailed model-selection procedure and selection results for the NHANES data. Then we present the analysis results of the NHANES data using the covariance regression model.

### 4.1. *Model selection for the NHANES data.*

Following the outline in Section 3.2, we first assume a constant covariance model and use AIC [Akaike (1973)] to select the set of predictors for the mean model. The best mean model under AIC includes the main effects of gender, age, race/ethnicity, and education, as well as four two-way interactions: gender and age, gender and race/ethnicity, gender and education, and age and race/ethnicity, with a total of 36 parameters. We then fix the mean model and select the explanatory variables in the covariance model. We first fit a rank-1 covariance regression model with main effects of the four predictors. We examine goodness of fit of this model and find that the observed values of the test statistics lie in the tails of the posterior predictive distributions for all groups. We then fit a rank-2 model with main effects of the four predictors. We plot the posterior predictive distributions of the test statistics in Figure 4. Three of the six subpopulations (gender and age, age and education, and race/ethnicity, and education) are well represented by the model. The remaining three subpopulations, which show lack of fit, all involve gender and/or race/ethnicity. Therefore, we add the interaction between gender and race/ethnicity to the rank-2 covariance model and present the goodness-of-fit diagnostics in Figure 5. The new model generally

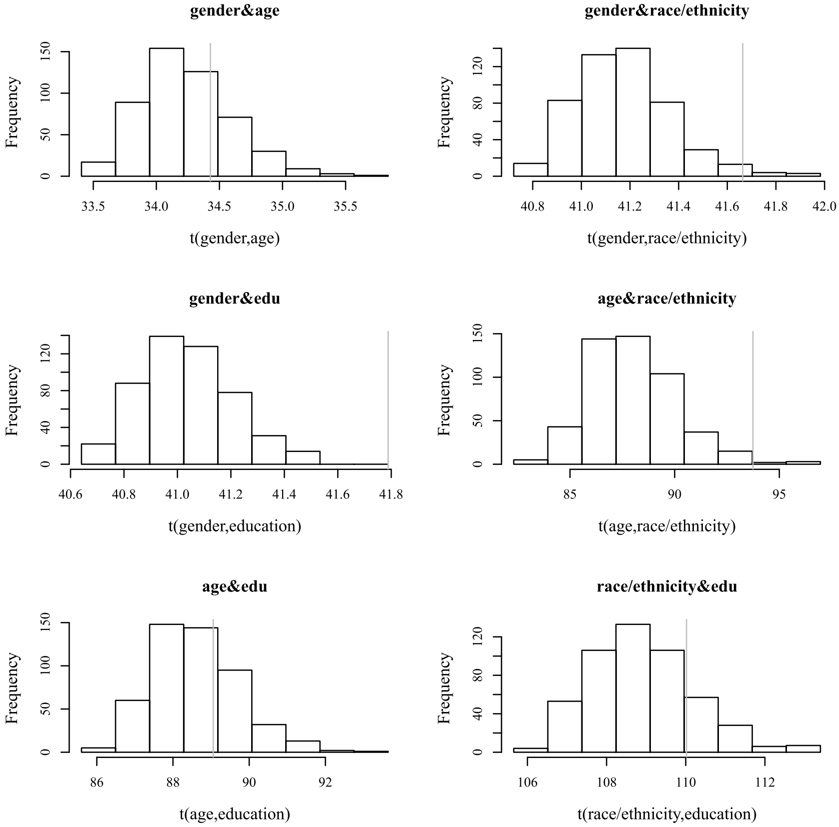

FIG. 4. *Posterior predictive distributions* (*of* 800 *posterior samples*) *of the rank-*2 *model with only main effects. In each histogram the vertical line represents the goodness-of-fit statistic calculated from the data.*

improves the goodness of fit from the rank-2 main-effects model, and it appropriately captures the heterogeneity in most of the 2-variable subpopulations. This relatively parsimonious model has no obvious lack of fit. We have also compared this model with adding other interaction terms and with a rank-3 model. None of those alternatives outperform this one. Therefore, our final model as stated in equation (3.5) and (3.6) includes the following terms in the mean and covariance structure:

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{x}_1} \sim {} & \mathrm{GENDER} + \mathrm{AGE} + \mathrm{RACE/ETHNICITY} + \mathrm{EDU} \\ & + \mathrm{GENDER} * \mathrm{AGE} + \mathrm{GENDER} * (\mathrm{RACE/ETHNICITY}) \\ & + \mathrm{GENDER} * \mathrm{EDU} + \mathrm{AGE} * (\mathrm{RACE/ETHNICITY}), \\ \boldsymbol{\Sigma}_{\mathbf{x}_2} \sim {} & \mathrm{GENDER} + \mathrm{AGE} + \mathrm{RACE/ETHNICITY} + \mathrm{EDU} \\ & + \mathrm{GENDER} * (\mathrm{RACE/ETHNICITY}).\end{aligned} \tag{4.1}$$

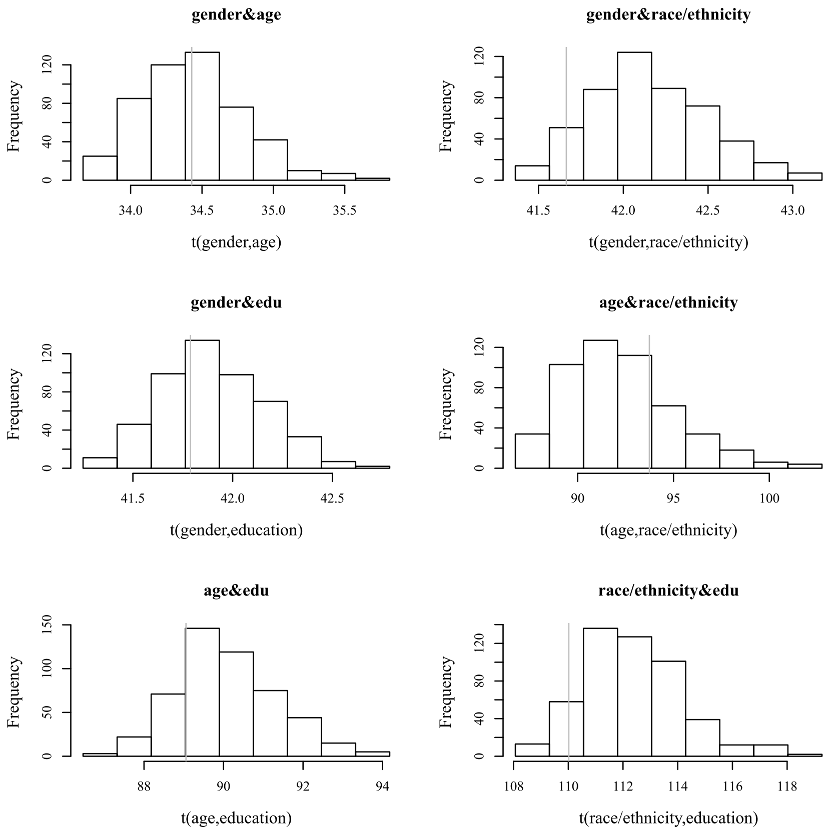

FIG. 5. *Posterior predictive distributions* (*of* 800 *posterior samples*) *of the rank-*2 *model with main effects and gender* $*$ (*race/ethnicity*) *interaction. In each histogram the vertical line indicates the goodness-of-fit statistic calculated from the data.*

4.2. *Results for NHANES data.* We fit the final model in equation (4.1) and obtain the Bayesian estimates through Gibbs sampling, using the priors described in Section 3.1. We run an MCMC chain for 50,000 iterations with thinning of every 50 samples (i.e., use every 50th iteration), drop the first 200 post-thinning samples as burn-in, and check the trace plots of key quantities for convergence. The analysis is performed using R-2.15.1, package "covreg". The key code to fit the model and summarize the results is in Supplementary Material C [Niu and Hoff (2019)]. The computation time is about 6 hours on a PC with an i5 core. It remains as future work to speed up the package in order to handle larger datasets.

There are multiple ways to present the fitted mean, variance, and correlation estimates of the four health measurements for all of the subgroups categorized by the four demographic characteristics. Here we suggest one graphic display that allows us to examine the relation of the posterior median estimates to a pair of demographic variables. For a scatter plot, we associate one category of the first variable

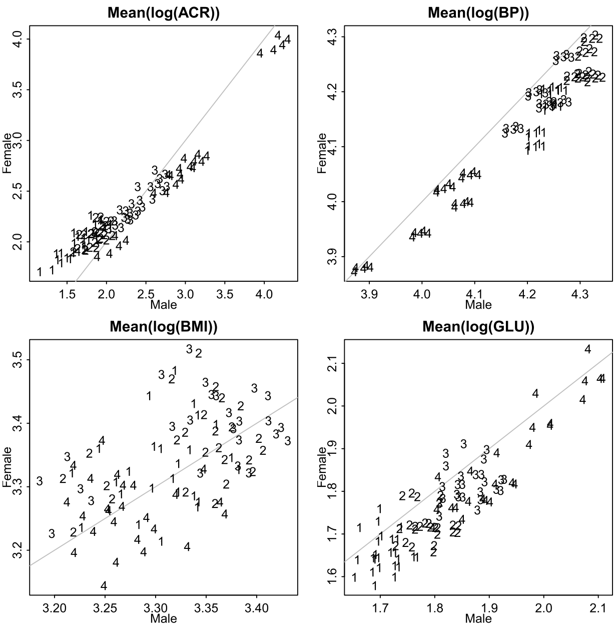

FIG. 6. *Scatter plots of group means by gender and age. For each combination of a category of age*, *a category of race/ethnicity*, *and a category of education*, *male and female define separate subpopulations in the* 4-*variable cross-classification*, *with the male group posterior median estimate as the $x$-coordinate and the female group posterior median estimate as the $y$-coordinate. The plotting symbols* 1 *to* 4 *represent the group's corresponding age category* (1: 20–39, 2: 40–59, 3: 60–79, 4: 80+). *The gray line is the reference line with slope* 1.

with the horizontal axis and a second category with the vertical axis. A digit corresponding to the category of the second variable serves as the plotting symbol. As an example, Figure 6 illustrates this basic structure with gender as the first variable and age as the second variable, with categories numbered 1 to 4. For each combination of a category of age, a category of race/ethnicity, and a category of education, male and female define separate subpopulations in the 4-variable cross-classification; the coordinates of the plotted point are the corresponding posterior median estimates.

These plots show how the mean, variance, and correlation vary with the demographic variables. We highlight a few interesting patterns. Figure 6 shows that females' ACR values are on average higher than the corresponding males' values groups in the younger age groups, but in the older age groups (some of the 60–79 groups and all of the 80+ groups) males' values are higher. Male groups' blood pressures are almost all higher than the values of the corresponding female groups. The 40–59 age groups have the highest average blood pressure. More male groups have higher glucose level than the corresponding female groups, and the glucose level seems to increase with age.

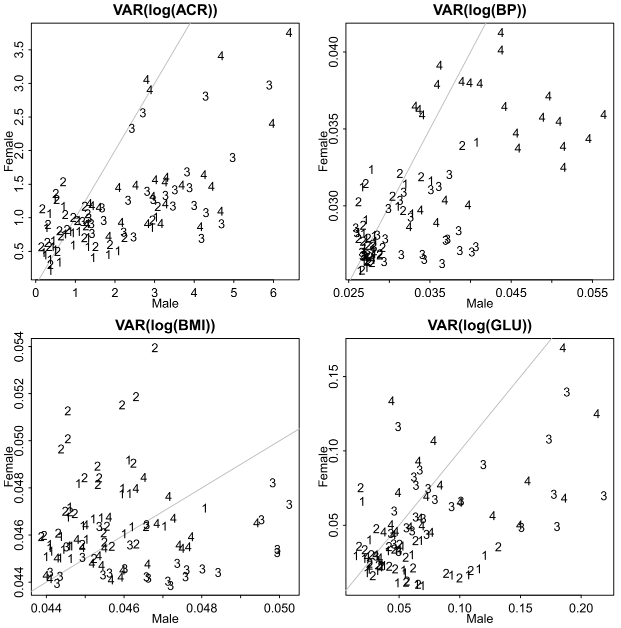

FIG. 7. *Scatter plots of group variances by gender and age. For each combination of a category of age, a category of race/ethnicity, and a category of education, male and female define separate subpopulations in the* 4*-variable cross-classification, with the male group posterior median estimate as the $x$-coordinate and the female group posterior median estimate as the $y$-coordinate. The plotting symbols* 1 *to* 4 *represent the group's corresponding age category* (1: 20–39, 2: 40–59, 3: 60–79, 4: 80+). *The gray line is the reference line with slope* 1.

In Figure 7, the variance of ACR and the variance of GLU vary greatly among subpopulations. The ratio of the largest to the smallest posterior median estimate of the standard deviation is 6.76 for ACR and 4.84 for GLU. On the other hand, the variance of BP and the variance of BMI do not vary too much; the ratios are 1.48 and 1.11, respectively. For ACR, most of the male group variance is larger than the corresponding female group, except for a few younger age groups. Variance generally seems to increase with age.

In the correlation plot of ACR and BMI (Figure 8), 60–79 year-old males have positive correlations, but the corresponding female groups have negative correlations. 40–59 year-old females have higher correlations than the corresponding male groups. We include similar plots of age and race/ethnicity and age and education in Supplementary Material D [Niu and Hoff (2019)]. In Supplementary Material E [Niu and Hoff (2019)] we present an alternative way of summarizing the results including estimation uncertainties.

The findings from our model provide some evidence that the mean, variance, and correlation of ACR, BP, BMI, and GLU vary among subpopulations. Therefore, it might not be appropriate to assume a common mean and variance for all

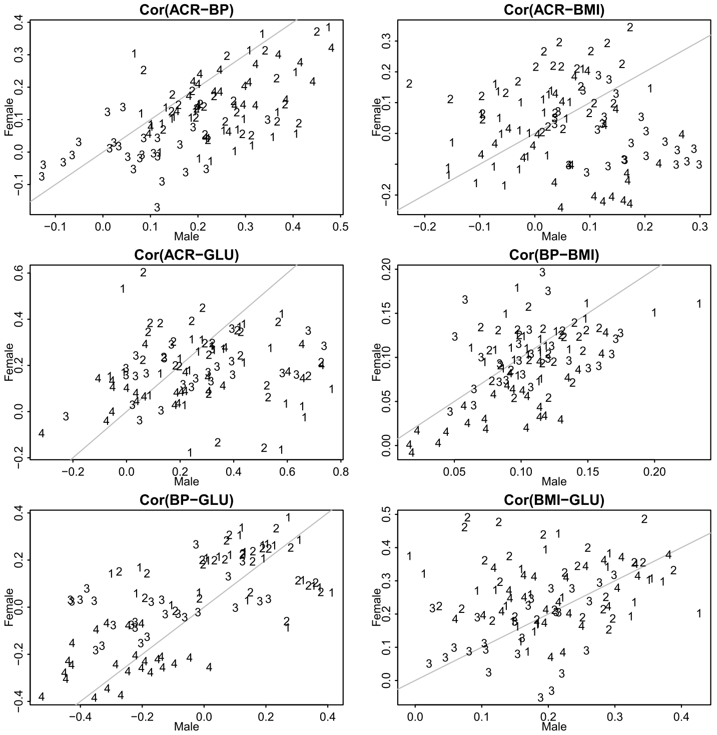

FIG. 8. *Scatter plots of group correlations by gender and age. For each combination of a category of age, a category of race/ethnicity, and a category of education, male and female define separate subpopulations in the* 4-*variable cross-classification, with the male group posterior median estimate as the $x$-coordinate and the female group posterior median estimate as the $y$-coordinate. The plotting symbols* 1 *to* 4 *represent the group's corresponding age category* (1: 20–39, 2: 40–59, 3: 60–79, 4: 80+). *The gray line is the reference line with slope* 1.

subpopulations when deriving reference ranges. More data need to be collected for some subpopulations to determine whether the reference ranges need to be refined.

**5. Discussion.** We use the NHANES data to study how the mean and covariance structure of four health measurements vary among subpopulations. To our knowledge, this is the first attempt to systematically examine how the variances and correlations of those health outcomes vary among subpopulations. We extend the covariance regression model proposed by Hoff and Niu (2012) to allow multiple categorical predictors, and we discuss practical issues in fitting complicated datasets with multiple categorical predictors. We select four highly relevant demographic and socio-economic factors to classify the population into subgroups and use the covariance regression model to estimate the mean and covariance parsimoniously for all subpopulations. We discuss guidelines for model selection and evaluation using standard criteria such as AIC for the mean model in conjunction with posterior predictive goodness-of-fit plots for the covariance model. The means, variances, and correlations of those health outcomes all vary among subgroups. The fitted results confirm that the population is heterogeneous and that assuming a single mean and a single covariance for the entire population is not appropriate. We highlight some of the findings that might be of scientific interest. The covariance regression model helps identify subpopulations for which more data might be collected to estimate a separate reference range. In Supplementary Material F [Niu and Hoff (2019)] we interpret some of the coefficient estimates.

We further validate the estimates by comparing the model-based intervals with the sample-based intervals for large groups (41 groups with sample size $> 20$). For the variance estimates, the percentage of times those two intervals overlap ranges from 85% to 95%, compared with 41% to 83% for the homogeneous model. The correlation estimates overlap 98% to 100%, compared with 85% to 95% for the homogeneous model. The complete set of plots comparing the two sets of intervals is in Supplementary Material G [Niu and Hoff (2019)]. To consider model misspecification, the sensitivity analysis [in Supplementary Material H (Niu and Hoff (2019)] gives some confidence that the covariance regression model is reasonable and provides reliable estimates.

The model-selection procedure can also allow the search to back up. If, after adding multiple interaction terms, some groups have over-fit, one can remove an interaction and refit, as in stepwise selection. The current selection and evaluation procedure is data-driven and subjective. To develop a systematic model-selection scheme that tries to find the overall best model, one possible approach could explicitly formulate a prior distribution to shrink some of the coefficients toward zero, similar to the idea in Gaskins and Daniels (2013).

NHANES uses a complex survey design to select samples that are representative of the U.S. noninstitutionalized civilian population. Adjusting the difference between sample and population is very important to obtain unbiased estimates of population quantities. As discussed in Gelman (2007), weighting and regression modeling are the two standard ways to accomplish this task. Winship and Radbill (1994) compare weighted and unweighted least-squares estimators for linear regression models and conclude that, if the weights depend on only predictors that are included in the model, and the model is true, (unweighted) OLS estimates are unbiased and consistent. In reality, if possible, accounting for the sampling weights is more accurate than using OLS estimates, because we never know whether the regression model is true, and often we cannot include all weight-determining factors and their interactions in the model [Gelman (2007)]. However, incorporating the sampling weights directly can be very difficult for nonstandard models. We therefore choose the regression approach by including the key factors that determine the weights in both the mean and covariance models. NHANES 2009–2010 oversampled specific age and race/ethnicity groups, as well as pregnant women [CDC/NCHS (2010b)]. Therefore, the key factors that determine the sampling weights are gender, age, and race/ethnicity, all of which we have included in the proposed joint model.

To further check for potential biases in our modelings, we compare the model-based estimates of the marginal cell means (such as mean ACR for all males) with the Horvitz–Thomson estimates of marginal cell means. We plot the H–T estimates, 95% confidence intervals for the H–T estimates, and model-based estimates in Supplementary Material I [Niu and Hoff (2019)]. The biggest discrepancy lies in the ACR estimates for males and females. This might be due to the fact that during pregnancy the ACR level might change, and we are not able to fully adjust for the oversampling of one gender (female) over the other (male). For most of the other groups, the two estimates are very close, and the model-based estimates almost all fall within the confidence intervals of the H–T estimates. This result gives us some confidence that our model provides approximately unbiased estimates without directly incorporating the sampling weights. It remains an interesting and challenging problem to directly incorporate the weights, thereby fully adjusting for the difference between sample and population.

In addition to the household survey, NHANES also selected certain participants for physical examinations, based on their demographic and health information. An even smaller proportion had blood tests. Of the 10,537 participants in the 2009–2010 survey, the numbers who had blood pressure measurements, urine tests, and body mass measurements range from 7000 to 9000. However, only 3386 participants had a fasting glucose value (blood test). Therefore, the main reduction in sample size is due to the design of the survey and can be viewed as a similar issue as weighting. On the other hand, individual measurements also have missing values (e.g., selected participants failed to show up for exams). Among those who have a GLU value, the missing proportions for the other three measurements (ACR, BP, BMI) are 1%, 4%, and 1%, respectively. According to the NHANES analysis guidelines (http://www.cdc.gov/nchs/tutorials/NHA-NES/Preparing/CleanRecode/Info1.htm), it is usually acceptable to ignore the missing values if the proportion is under 10%. Therefore, after accounting for the design variables as discussed above, we assume the missingness due to nonresponse is ignorable, and we use the complete data for the analysis without any imputation.

The current model assumes multivariate normality of the error term, which is a strong assumption. We assessed some residual plots and did not see serious violations. In essence, we are trying to model the first- and second- order moments that do not rely on normality of the data. The model could be extended to nonnormally distributed variables through the generalized linear model framework, as in Pourahmadi (1999). Another possibility is via semi-parametric copula models, as proposed by Hoff (2007).

In this study, we focus on four specific health outcomes. The method is general enough to be applied to a wide variety of multivariate outcomes.

**Acknowledgements.** The authors thank the reviewers, Associate Editor, and Editor for their helpful comments. We are especially grateful for the Associate Editor's careful review and edits that led to a stronger article.

## SUPPLEMENTARY MATERIAL

**Supplement to "Joint mean and covariance modeling of multiple health outcome measures."** (DOI: 10.1214/18-AOAS1187SUPP; .pdf). Additional results, tables, and plots mentioned in the text are in the Supplemental Material.

## REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) (B. N. Petrov and F. Csaki, eds.) 267–281. Akadémiai Kiadó, Budapest. MR0483125

BOIK, R. J. (2002). Spectral models for covariance matrices. *Biometrika* **89** 159–182. MR1888370

BOIK, R. J. (2003). Principal component models for correlation matrices. *Biometrika* **90** 679–701. MR2006844

CDC/NCHS (2010a). National Health and Nutrition Examination Survey Data, 2009–2010. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD.

CDC/NCHS (2010b). National Health and Nutrition Examination Survey: Analytic Guidelines, 1999–2010. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD.

CHIU, T. Y. M., LEONARD, T. and TSUI, K.-W. (1996). The matrix-logarithmic covariance model. *J. Amer. Statist. Assoc.* **91** 198–210. MR1394074

CLSI (2008). *Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory: Approved Guideline*, 3rd ed. CLSI document EP28-A3c. Clinical and Laboratory Standards Institute, Wayne, PA.

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B* **49** 1–39 (with a discussion). MR0893334

CRIPPS, E., CARTER, C. and KOHN, R. (2005). Variable selection and covariance selection in multivariate regression models. In *Bayesian Thinking: Modeling and Computation* (D. Dey and C. R. Rao, eds.). *Handbook of Statist.* **25** 519–552. Elsevier/North-Holland, Amsterdam. MR2490538

ENGLE, R. F. and KRONER, K. F. (1995). Multivariate simultaneous generalized ARCH. *Econometric Theory* **11** 122–150. MR1325104

FONG, P. W., LI, W. K. and AN, H.-Z. (2006). A simple multivariate ARCH model specified by random coefficients. *Comput. Statist. Data Anal.* **51** 1779–1802. MR2307543

FOULDS, H., BREDIN, S. and WARBURTON, D. (2012). The relationship between diabetes and obesity across different ethnicities. *J. Diabetes Metab.* **3**.

FRASER, S. D. S., RODERICK, P. J., MCLNTYRE, N. J., HARRIS, S., MCLNTYRE, C. W., FLUCK, R. J. and TAAL, M. W. (2012). Socio-economic disparities in the distribution of cardiovascular risk in chronic kidney disease stage 3. *Nephron, Clin. Pract.* **122** 58–65.

GASKINS, J. T. and DANIELS, M. J. (2013). A nonparametric prior for simultaneous covariance estimation. *Biometrika* **100** 125–138. MR3034328

GELMAN, A. (2007). Struggles with survey weighting and regression modeling. *Statist. Sci.* **22** 153–164. MR2408951

GUTTMAN, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *J. Roy. Statist. Soc. Ser. B* **29** 83–100. MR0216699

HARRIS, E. K. and BOYD, J. C. (1990). On dividing reference data into subgroups to produce separate reference ranges. *Clin. Chem.* **36** 265–270.

HOFF, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.* **1** 265–283. MR2393851

HOFF, P. D. (2009). A hierarchical eigenmodel for pooled covariance estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 971–992. MR2750253

HOFF, P. D. and NIU, X. (2012). A covariance regression model. *Statist. Sinica* **22** 729–753. MR2954359

KDIGO (2013). Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Inter., Suppl.* **3** 1–150.

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

MATTIX, H. J., HSU, C.-Y., SHAYKEVICH, S. and CURHAN, G. (2002). Use of the albumin/creatinine ratio to detect microalbuminuria: Implications of sex and race. *J. Am. Soc. Nephrol.* **13** 1034–1039.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, London. [Second edition of MR0727836.] MR3223057

NIDDK (2013). U.S. Renal Data System, USRDS 2013 Annual Data Report: Atlas of Chronic Kidney Disease and End-Stage Renal Disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD.

NIU, X. and HOFF, P. D. (2019). Supplement to "Joint mean and covariance modeling of multiple health outcome measures." DOI:10.1214/18-AOAS1187SUPP.

POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. MR1723786

POURAHMADI, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statist. Sci.* **26** 369–387. MR2917961

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

WINSHIP, C. and RADBILL, L. (1994). Sampling weights and regression analysis. *Sociol. Methods Res.* **23** 230–257.

ZEGER, S. L. and LIANG, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 121–130.

DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
323C THOMAS BUILDING
UNIVERSITY PARK, PENNSYLVANIA 16802
USA
E-MAIL: xiaoyue@psu.edu

DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
219 OLD CHEMISTRY BUILDING
BOX 90251
DURHAM, NORTH CAROLINA 27708-0251
USA
E-MAIL: peter.hoff@duke.edu