DISCUSSION PAPER SERIES

Discussion paper No. 177

Bayesian factor models for probabilistic cause of death assessment with verbal autopsies

Tsuyoshi Kunihama
(Kwansei Gakuin University)
Zehang Richard Li
(University of Washington)
Samuel J. Clark
(Ohio State University)
Tyler H. McCormick
(University of Washington)

March 2018

SCHOOL OF ECONOMICS

KWANSEI GAKUIN UNIVERSITY

1-155 Uegahara Ichiban-cho
Nishinomiya 662-8501, Japan

Bayesian factor models for probabilistic cause of death assessment with verbal autopsies

Tsuyoshi Kunihama
Kwansei Gakuin University

Zehang Richard Li
University of Washington

Samuel J. Clark
Ohio State University

Tyler H. McCormick*
University of Washington

Abstract

The distribution of deaths by cause provides crucial information for public health planning, response, and evaluation. About 60% of deaths globally are not registered or given a cause which limits our ability to understand the epidemiology of affected populations. Verbal autopsy (VA) surveys are increasingly used in such settings to collect information on the signs, symptoms, and medical history of people who have recently died. This article develops a novel Bayesian method for estimation of population distributions of deaths by cause using verbal autopsy data. The proposed approach is based on a multivariate probit model where associations among items in questionnaires are flexibly induced by latent factors. We measure strength of conditional dependence of symptoms with causes. Using the Population Health Metrics Research Consortium labeled data that include both VA and medically certified causes of death, we assess performance of the proposed method. Further, we propose a method to estimate important questionnaire items that are highly associated with causes of death. This framework provides insights that will simplify future data collection.

Key words: Bayesian latent model; Cause of death; Conditional dependence; Multivariate data; Verbal autopsies; Survey data.

1 Introduction

Data on cause of death are essential to understand the epidemiology of a population, to design and implement efficient Public Health interventions, and to measure their effects (e.g. Ruzicka and Lopez, 1990; Mathers et al., 2005; Soleman et al., 2006; Bloomberg and Bishop, 2015). Understanding progress in controlling an infection disease, for example, requires monitoring changes in the cause of death distribution in populations over time. The "gold-standard" for assigning cause of death relies on physical autopsies with pathological

*Correspondence contact tylermc@uw.edu

reports, backed by a comprehensive registration system. In low-resource settings, however, most deaths happen outside of hospitals and are often not recorded by a civil registration and vital statistics systems (Mikkelsen et al., 2015). In such settings, understanding the mortality burden of a specific cause, and trends in cause-specific mortality over time, is extremely challenging (e.g. AbouZahr et al., 2007; Boerma and Stansfield, 2007; Hill et al., 2007; Mahapatra et al., 2007; Setel et al., 2007; Phillips et al., 2014; de Savigny et al., 2017; Phillips et al., 2015).

Scaling up to a full-coverage civil registration system presents massive financial and logistical challenges, meaning that survey-based data are and will continue to be vital for understanding cause of death distributions (Horton, 2007; AbouZahr et al., 2007; Jha, 2014). These survey-based data, known as a verbal autopsies (VAs), consist of interviews with a family member or other individual familiar with the death. The respondent answers a questionnaire about the signs, symptoms, demographic characteristics and health history of the deceased. Deaths are typically identified using community informants or using a partial surveillance system. VA surveys are widely conducted (Lopez, 1998; Yang et al., 2005; Maher et al., 2010; Sankoh and Byass, 2012) and the World Health Organization (WHO) releases a the standardized VA questionnaire (Baiden et al., 2007; World Health Organization, 2012, 2017; Nichols et al., 2018) to facilitate comparison across areas.

VA surveys are substantially more cost effective than performing in-person autopsies. Unlike a physical autopsy, however, VA surveys require an additional step to assign a cause of death from the collected symptoms. Many methods have been proposed to estimate causes from VA interview data. In some settings, trained clinicians review VAs and assess a cause of death (Lozano et al., 2011). This approach can be effective in some circumstances, but is time-consuming and requires that trained clinicians (many of whom would otherwise be seeing patients) be available. An alternative approach is to use an algorithmic or statistical method to assign causes of death. Several such methods have been proposed and evaluated in the statistics and public health literatures (see for example Murray et al., 2007; King and Lu, 2008; James et al., 2011; King et al., 2010; Murray et al., 2011; Byass et al., 2012; Serina et al., 2015; Miasnikof et al., 2015; McCormick et al., 2016).

For the most part, these methods rely on a critical assumption: independence across

symptoms conditional on a given cause. This assumption disregards critical information about constellations or clusters of symptoms that are typical of a given cause and thus particularly informative when assigning causes. The only method currently available that uses information about dependence between symptoms is work by King and Lu (King and Lu, 2008; King et al., 2010). The King and Lu method regresses the cause of death on randomly sampled sets of symptoms from a training dataset. This process is an attempt to represent the space of all possible symptom combinations. However since there are typically one to two hundred symptoms, exploring all possible combinations is a daunting task, or practically impossible.

Our work presents a novel approach to incorporating dependence between symptoms in coding cause of death from VA surveys. In our approach, we capture dependence between symptoms using a small number of latent factors. This approach avoids the need to evaluate all possible symptom combinations as in the King and Lu framework. We build a multivariate probit model for symptoms conditional on a cause. Binary-scale outcomes can be interpreted as a manifestation of underlying continuous variables. A factor model on these conditional variables provides a sparse covariance structure between symptoms. Our method also accommodates missing data that commonly arise in VA surveys, because for example, family members may not remember all details about sign/symptoms of the deceased. The proposed approach can incorporate both individual-specific and design-based missing values by summing them out from the probit model with a missing-at-random assumption. We fit the model using an efficient Markov chain Monte Carlo (MCMC) algorithm we develop for posterior computation. Further, we utilize our framework to better understand the importance of each measure in the questionnaire. To do this, we quantify the association between each symptom with each cause given all other predictors, using conditional mutual information to measure association. Our measures can be used to simplify and shorten future VA surveys, decreasing both the burden on respondents and the cost.

The rest of the paper is organized as follows. The remainder of this section describes labeled VA data from the Population Health Metrics Research Consortium that we will use in our analysis. Section 2 proposes a novel approach for estimation of population distributions of causes of death. Section 3 develops an efficient MCMC algorithm for the

proposed method. Section 4 assesses the performance of the proposed approach in various scenarios and measures strength of conditional dependence of questionnaire items in the gold-standard dataset. Section 5 concludes the article.

1.1 PHMRC VA survey

The Population Health Metrics Research Consortium (PHMRC) collected VA data at six study sites in four countries: Andhra Pradesh, India (AP); Bohol, Philippines (Bohol); Dar es Salaam, Tanzania (Dar); Mexico City, Mexico (Mexico); Pemba Island, Tanzania (Pemba); and Uttar Pradesh, India (UP). In each study site, VAs were collected for adults, children and neonates in hospital and clinical environments. Causes were assigned based on diagnostic criteria including laboratory, pathology and medical imaging findings (Murray et al., 2011). VA interviews were conducted with a relative of the deceased by interviewers who were blinded to the cause of death assigned in the hospital. The VA questionnaire items cover symptoms of illnesses, demographic characteristics, diagnoses of chronic illnesses by health service providers, possible risk factors such as tobacco-use and other potentially contributing characteristics. In typical settings where VA surveys are implemented, it is not possible to obtain a large fraction of deaths with physician codes. The PHMRC data are, therefore, a "gold-standard" dataset for training VA methods to be applied in situations where physician coding is not possible. PHMRC (2013) distributes the data and the codebook to the public.

Figure 1 shows histograms of 34 causes of death for adults in the six study sites, indicating that distributions of the causes vary considerably among the sites. The VA questionnaire consists of binary, count and categorical items. As is the standard for analyzing VA data, we pre-process the mixed-scale questions into a combination of binary variables, and assume that all items are dichotomous using the same steps as described in McCormick et al. (2016), leading to the data set with 7,841 individuals and 175 items. Also, there are abundant missing values in the data because respondents do not remember all details about symptoms of the deceased. Figure 2 shows boxplots of frequencies of answering yes and missing rates for the binary questions. The medians are small but we observe several outliers in both cases, indicating that some questions are missing in nearly all of the cases.

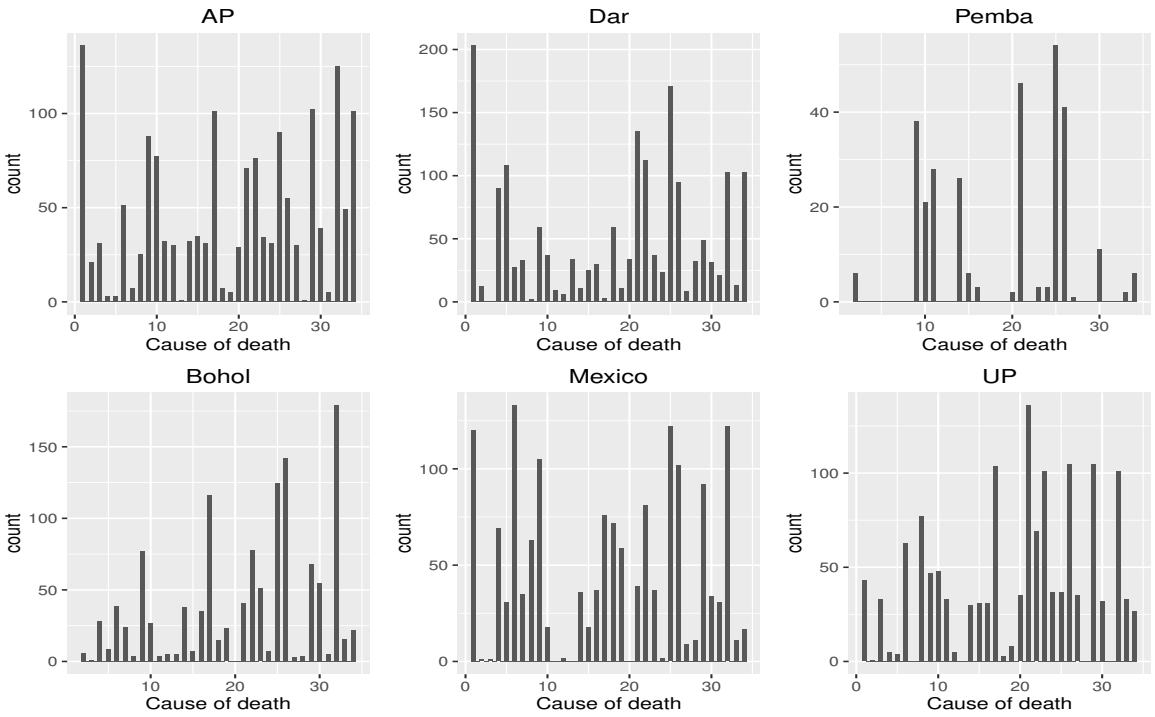

Figure 1: Histograms of causes of death for the PHMRC study sites.

Since the PHMRC data contain medically-certified causes, we can explore the magnitude of the dependence between symptoms for a given medically-certified cause. We compute Cramér's V that measures strength of associations between two variables, taking a value from 0 (no association) to 1 (complete association). Figure 3 shows the result for all pairs of symptoms for the deaths caused by AIDS and Stroke. We conducted chi-squared test using the R function cramersV in the lsr package (Navarro, 2015; R Core Team, 2016), and the hypothesis of independence was rejected with 5% significance level for more than 1,000 pairs of the symptoms. Unlike nearly all previously available methods, our proposed method will utilize these correlations to improve cause assignment accuracy.

2 Bayesian factor model for VA data

In this section we present our model formulation. First we present the modeling framework, and then we discuss how we can compute conditional mutual information as a means of assessing the importance of symptoms.

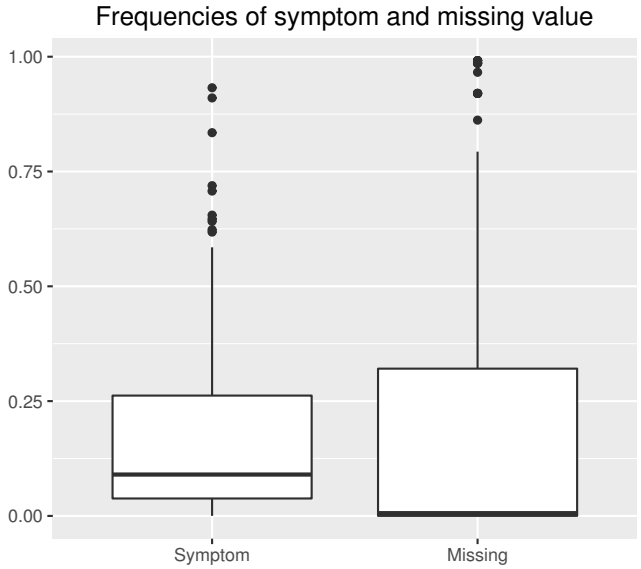

Figure 2: Boxplots of frequencies of symptoms and missing values.

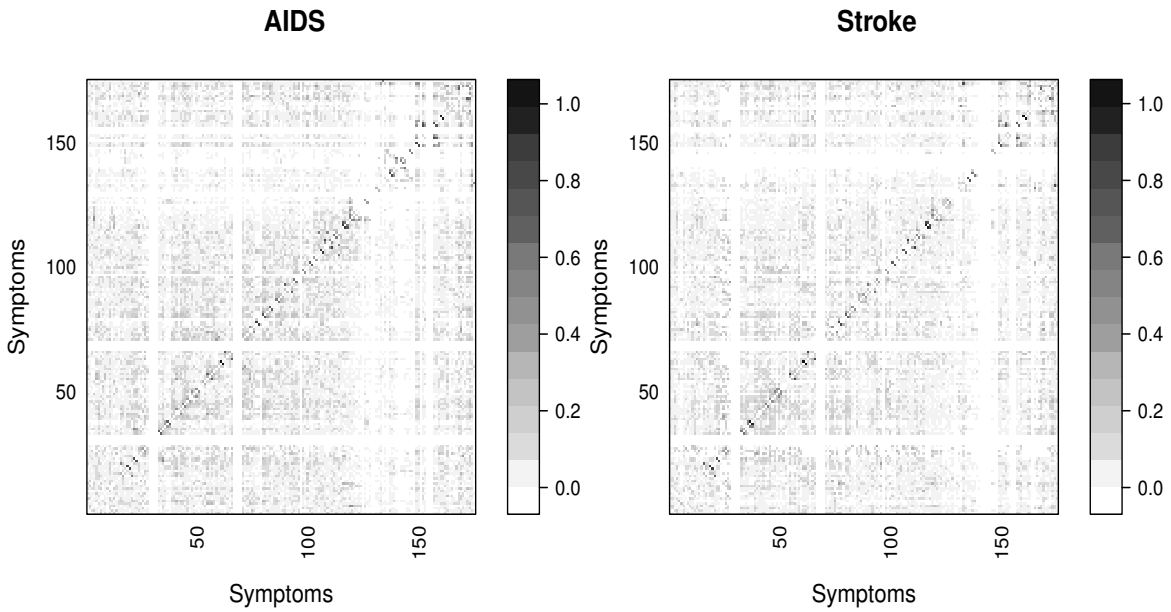

Figure 3: Cramér's V among symptoms for AIDS (left) and Stroke (right).

2.1 Bayesian approach

We propose a novel Bayesian framework for assessing cause of death using VA surveys. Let $y_i \in \{1, \ldots, C\}$ be the cause responsible for ith person's death with $i = 1, \ldots, n$, and $x_i = (x_{i1}, \ldots, x_{ip})'$ be the response to questions $j = 1, \ldots, p$ with $x_{ij} \in \{0, 1\}$. One approach would be to directly build a conditional probability $\pi(y_i \,|\, x_i)$ relying on standard parametric models such as multinomial probit/logit regressions. However as discussed in the previous section, modeling these conditional probabilities directly is unappealing since we would need to impute a substantial fraction of symptoms (see Figure 2). Further, we would need to impute these symptoms using very little information from the data, meaning that the choice of imputation method could be influential in our cause of death assignment.

Instead, we will express the conditional distribution, $\pi(y_i \,|\, x_i)$, using Bayes' rule, a choice that facilitates directly integrating over the missing data. To see this, take first

$$\pi(y_i \,|\, x_i) = \frac{\pi(x_i \,|\, y_i)\pi(y_i)}{\sum_{c=1}^{C} \pi(x_i \,|\, y_i = c)\pi(y_i = c)}.$$

In this framework we can incorporate missing values easily by integrating them out from $\pi(x_i \,|\, y_i)$. Let x_i^{obs} and x_i^{mis} denote the observed and missing items for the ith person with $x_i = (x_i^{obs}, x_i^{mis})$. Assuming symptoms are missing at random, we utilize the distribution of the cause given the observed items,

$$\pi(y_i \,|\, x_i^{obs}) \propto \pi(x_i^{obs} \,|\, y_i)\pi(y_i), \text{ where } \pi(x_i^{obs} \,|\, y_i) = \int \pi(x_i \,|\, y_i) dx_i^{mis}.$$

Therefore if we can calculate the integral analytically, we can evaluate the conditional probabilities of the cause on the observed information without imputing missing data.

Since we are using a Bayes' rule representation, there are two pieces of the model that we need to specify, (i) the unconditional distribution of individual causes, $\pi(y_i)$, and (ii) the conditional distribution of observed symptoms given an individual has a particular cause, $\pi(x_i \,|\, y_i)$. Beginning with the prior distribution for causes, we assume a Dirichlet distribution,

$$\{\pi(y_i = 1), \ldots, \pi(y_i = C)\} \sim \text{Dirichlet}(a_1, \ldots, a_C)$$

where $a_1, \ldots, a_C$ are concentration parameters. Since cause patterns differ substantially across geographic areas and time, we assume we have little prior information about the distribution of causes. We therefore assume $a_1 = \cdots = a_C = 1$, leading to a uniform prior with $\pi(y_i = c) \propto 1$ for $c = 1, \ldots, C$.

The second piece, $\pi(x_i \,|\, y_i)$, requires modeling a set of high-dimensional binary symptoms given each cause. As described previously, nearly all existing methods for assigning cause of death from verbal autopsies make a conditional independence assumption across symptoms (Byass et al., 2012; Miasnikof et al., 2015; McCormick et al., 2016),

$$\pi(x_i \,|\, y_i) = \prod_{j=1}^{p} \pi(x_{ij} \,|\, y_i).$$

This assumption facilitates computation but disregards substantial and potentially informative information about the relationships between symptoms, as Figure 3 shows.

To flexibly capture dependence, we develop a conditional distribution based on the multivariate probit model. In our framework, each binary outcome is a manifestation of an underlying continuous variable. Let $z_i = (z_{i1}, \ldots, z_{ip})' \in \mathbb{R}^p$ be the latent variable for ith person, inducing the symptoms via an indicator function. We assume a multivariate normal distribution conditional on a cause, $z_i \,|\, y_i \sim N(\mu_{y_i}, \Sigma_{y_i})$ with mean $\mu_{y_i} = (\mu_{y_i 1}, \ldots, \mu_{y_i p})'$ and covariance Σ_{y_i}. There are $p(p+1)/2$ parameters in the covariance.

Even for moderately large p, estimating this many parameters will be challenging, particularly since we expect that each dataset will contain only a few deaths by each cause. Further, since our goal is predicting cause of death for a new sample of deaths, we prefer a sparse model to minimize issues with generalization arising from overfitting. Rather than estimating all elements in the covariance matrix, we introduce a K-dimensional factor $\eta_i = (\eta_{i1}, \ldots, \eta_{iK})'$ with $K \ll p$, and propose the following sparse factor model,

$$\begin{aligned} x_{ij} &= 1(z_{ij} > 0), \quad j = 1, \ldots, p, \\ z_i &= \mu_{y_i} + \Lambda_{y_i} \eta_i + \varepsilon_i, \quad \eta_i \sim N(0, I_K), \quad \varepsilon_i \sim N(0, I_p), \end{aligned} \tag{1}$$

where $1(\cdot)$ is an indicator function and $\Lambda_y = \{\lambda_{yjk}\}$ is a $p \times K$ loading matrix with $y = 1, \ldots, C$, $j = 1, \ldots, p$ and $k = 1, \ldots, K$. Dependence is induced in z_i by integrating out

the factor η_i in (1), leading to the normal distribution with the cause-dependent mean and covariance,

$$z_i \,|\, y_i \sim N(\mu_{y_i}, \Lambda_{y_i}\Lambda'_{y_i} + I_p),$$

where the number of parameters in the covariance reduces from $p(p+1)/2$ to Kp. For the prior distribution of the mean and factor loadings, we use Cauchy distributions, a standard shrinkage prior with high density around zero and heavy tails, reducing effects of redundant elements but capturing important signals. Based on the convolution expression of the Cauchy distribution, we assume

$$\begin{aligned} \mu_{yj} &\sim N(0, \tau_j^{-1}), \quad \tau_j \sim Ga(0.5, 0.5), \\ \lambda_{yjk} &\sim N(0, \phi_j^{-1}), \quad \phi_j \sim Ga(0.5, 0.5), \end{aligned}$$

where $Ga(a, b)$ denotes the Gamma distribution with mean a/b. The latent variables τ_j and ϕ_j are shared among the causes and factors for reduction of parameters in the model.

2.2 Measuring strength of conditional associations

We now present a method to ascertain the information that each symptom provides in addition to other symptoms already in the model. We present the association metric in this section and then, in the following section, describe how the measure can be incorporated into our posterior sampling algorithm.

VA surveys collect information via questionnaires with many items regarding demographic background, health history and disease symptoms. It can be time-consuming and costly to ask redundant questions with little information for prediction of causes of death. Further, the VA interview requires that a family member or other person close to the decedent recall a potentially painful and traumatic time. Consequently, asking questions that do not inform cause of death classification prolongs this potentially negative experience for the respondent with no benefits to public health and potential cost implications for the survey as a whole.

As a measure of strength of conditional dependence, we compute conditional mutual

information, a quantity defined in the information theory literature. Let ζ_j denote the conditional mutual information for the jth predictor. Conditional mutual information quantifies the change in the distribution of $\pi(y \mid x)$ associated with adding the j predictor, conditional on all other predictors being added already. Large values for conditional mutual information indicate strong associations of the jth item with the response. Conditional mutual information is non-negative and equals zero if and only if the information in the jth item is redundant conditional on other predictors (Wyner, 1978; Joe, 1989; Cover and Thomas, 2006). Formally, ζ_j can be defined by the expectation of the Kullback-Leibler divergence between $\pi(y \mid x)$ and $\pi(y \mid x_{-j})$ with $x_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)'$,

$$\zeta_j = E\left[KL\{\pi(y \mid x), \pi(y \mid x_{-j})\}\right] = \sum_{y=1}^{C} \sum_{x \in \mathcal{X}} \pi(y, x) \log \frac{\pi(y \mid x)}{\pi(y \mid x_{-j})},$$

where $\mathcal{X} = \{0, 1\}^p$. Based on the Bayes' theorem, this can be expressed as

$$\zeta_j = \sum_{y=1}^{C} \sum_{x \in \mathcal{X}} \pi(y, x) \log \frac{\pi(x \mid y)\pi(x_{-j})}{\pi(x_{-j} \mid y)\pi(x)}.$$

It is computationally intractable to evaluate the summation over the space $\mathcal{X}$ for large p. As a solution we apply the Monte Carlo approximation using observations (Kunihama and Dunson, 2016). Recall that the standard Monte Carlo approximation in this setting would be

$$\zeta_j = \sum_{y=1}^{C} \sum_{x \in \mathcal{X}} \pi(y, x) \log \frac{\pi(x \mid y)\pi(x_{-j})}{\pi(x_{-j} \mid y)\pi(x)} \approx \frac{1}{R} \sum_{r=1}^{R} \log \frac{\pi(x_r \mid y_r)\pi(x_{r,-j})}{\pi(x_{r,-j} \mid y_r)\pi(x_r)},$$

where $\{(y_r, x_r),\ r = 1, \ldots, R\}$ is the random sample from the proposed model $\pi(y_i, x_i)$. This approach would still be computationally burdensome, as it would require sampling R observations from the proposed model and evaluating the above expression at each step of

the sampler. If the proposed model approximates the true one well, then

$$\begin{aligned}\zeta_j &= \sum_{y=1}^{C}\sum_{x\in\mathcal{X}} \pi(y,x)\log\frac{\pi(x\,|\,y)\pi(x_{-j})}{\pi(x_{-j}\,|\,y)\pi(x)} \approx \sum_{y=1}^{C}\sum_{x\in\mathcal{X}} \pi_0(y,x)\log\frac{\pi(x\,|\,y)\pi(x_{-j})}{\pi(x_{-j}\,|\,y)\pi(x)},\\ &\approx \frac{1}{n}\sum_{i=1}^{n}\log\frac{\pi(x_i\,|\,y_i)\pi(x_{i,-j})}{\pi(x_{i,-j}\,|\,y_i)\pi(x_i)}, \end{aligned} \tag{2}$$

where $\{(y_i, x_i),\ i = 1,\ldots,n\}$ is the random sample from the true $\pi_0(y,x)$, that is, the observations. Under the assumption that the true model is close to the proposed model, using the observations means that we can avoid generating the large Monte Carlo sample $\{(y_r, x_r),\ r = 1,\ldots,R\}$ at each MCMC iteration.

For the proposed model, it is straightforward to evaluate the probability functions in (2). The $\pi(x_i)$ term, for example, can be computed as a simple summation over causes, $\pi(x_i) = \sum_{y=1}^{C}\pi(x_i\,|\,y)\pi(y)$. If there are many uninformative symptoms (and thus a sum with multiple terms that are close to log(1)), the approximation may break the non-negativity constraint on ζ_j. However for questions highly associated with causes given the other items, ζ_j will be far away from zero. Alternatively, the approximation with observations can be viewed as a comparison of data-fitness between $\pi(y\,|\,x)$ and $\pi(y\,|\,x_{-j})$. A positive value indicates that the fitness of data improves by adding the jth item, while a negative one implies that the additional predictor x_j causes a gap between the model and data. For example, if x_j highly correlates with x_{-j} but has no additional information related to y, then it induces just complexity in $\pi(x\,|\,y)$ for $y = 1,\ldots,C$, leading to the poor data-fit of $\pi(y\,|\,x) \propto \pi(x\,|\,y)\pi(y)$ compared to $\pi(y\,|\,x_{-j})$.

3 Posterior computation

The posterior density for the model presented in Section 2.1 is not available in closed form. We instead approximate the posterior density using samples obtained through Markov-chain Monte Carlo (MCMC). Let $m_i = (m_{i1},\ldots,m_{ip})'$ be a vector of indicators denoting missing values for the ith person such that $m_{ij} = 1$ if x_{ij} is missing and $m_{ij} = 0$ if x_{ij} is observed with $j = 1,\ldots,p$. We define notation $[m_i]$ such that, for a vector b and a matrix B with p rows, $b_{[m_i]}$ and $B_{[m_i]}$ denote the subvector and submatrix consisting of components with

$m_{ij} = 0$ for $j = 1, \ldots, p$. Then we propose the following MCMC algorithm.

1. Update $\mu_{\cdot j} \equiv (\mu_{1j}, \ldots, \mu_{Cj})'$ from $N(\mu_*, \Sigma_*)$ for $j = 1, \ldots, p$ with

$$\mu_* = \Sigma_* a_j, \quad \Sigma_* = \mathrm{diag}\left\{(n_1 + \tau_j)^{-1}, \ldots, (n_C + \tau_j)^{-1}\right\},$$

   where $n_c = \sum_{i=1}^n 1(y_i = c, m_i = 0)$ and a_j is the $C \times 1$ vector with the cth element $\sum_{i=1}^n 1(y_i = c, m_i = 0)(z_{ij} - \lambda'_{y_i j \cdot}\eta_i)$ where $\lambda_{y_i j \cdot} = (\lambda_{y_i j 1}, \ldots, \lambda_{y_i j K})'$.

2. Update $\lambda_{cj \cdot} \equiv (\lambda_{cj1}, \ldots, \lambda_{cjK})'$ from $N(\mu_\lambda, \Sigma_\lambda)$ for $c = 1, \ldots, C$ with

$$\mu_\lambda = \Sigma_\lambda \left\{ \sum_{i:y_i=c} \eta_i (z_{ij} - \mu_{y_i j}) \right\}, \quad \Sigma_\lambda = \left(\sum_{i:y_i=c} \eta_i \eta_i' + \phi_j I_K \right)^{-1}.$$

3. Update η_i from $N(\tilde{\mu}, \tilde{\Sigma})$ for $i = 1, \ldots, n$ with

$$\tilde{\mu} = \tilde{\Sigma} \Lambda_{y_i[m_i]} (z_i - \mu_{y_i})_{[m_i]}, \quad \tilde{\Sigma} = \left(\Lambda'_{y_i[m_i]} \Lambda_{y_i[m_i]} + I_K \right)^{-1}.$$

4. Update τ_j for $j = 1, \ldots, p$ from

$$Ga\left(\frac{C+1}{2}, \frac{\sum_{c=1}^C \mu_{cj}^2 + 1}{2} \right).$$

5. Update ϕ_j for $j = 1, \ldots, p$ from

$$Ga\left(\frac{CK+1}{2}, \frac{\sum_{c=1}^C \sum_{k=1}^K \lambda_{cjk}^2 + 1}{2} \right).$$

6. Update z_{ij} with $m_{ij} = 0$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$ from

$$\begin{cases} N_+(\mu_{y_i j} + \lambda'_{y_i j \cdot}\eta_i, 1) & \text{if } x_{ij} = 1, \\ N_-(\mu_{y_i j} + \lambda'_{y_i j \cdot}\eta_i, 1) & \text{if } x_{ij} = 0, \end{cases}$$

   where N_+ and N_- denote the truncated normal distributions with support $[0, \infty)$ and $(-\infty, 0]$ respectively.

7. For a person $i \in S$ where S is the target data, generate y_i with

$$\pi(y_i = c \,|\, x_i^{obs}) = \frac{\pi(x_i^{obs} \,|\, y_i = c)\pi(y_i = c)}{\sum_{y=1}^{C} \pi(x_i^{obs} \,|\, y_i = y)\pi(y_i = y)}, \quad c = 1, \ldots, C,$$

where $\pi(x_i^{obs} \,|\, y_i = c) = \int \pi(x_i^{obs} \,|\, \eta, y_i = c) f(\eta) d\eta$ is evaluated using a Monte Carlo approximation with $\eta_r \sim N(0, I_K)$ for $r = 1, \ldots, R$,

$$\pi(x_i^{obs} \,|\, y_i = c) \approx \frac{1}{R} \sum_{r=1}^{R} \pi(x_i^{obs} \,|\, \eta_r, y_i = c) = \frac{1}{R} \sum_{r=1}^{R} \left\{ \prod_{j: m_{ij}=0} \pi(x_{ij} \,|\, \eta_r, y_i = c) \right\}. \quad (3)$$

Then, compute the population distribution of causes of death by

$$\left(\frac{1}{\#S} \sum_{i \in S} 1(y_i = 1), \ldots, \frac{1}{\#S} \sum_{i \in S} 1(y_i = C) \right)$$

where $\#S$ is the number of observations in the target data.

For the estimation of strength of conditional dependence in Section 3.2, Step 7 above is replaced by

7. Update the distribution of causes with the prior Dirichlet$(1, \ldots, 1)$ from

$$\{\pi(y_i = 1), \ldots, \pi(y_i = C)\} \sim \text{Dirichlet} \left(\frac{1}{n} \sum_{i=1}^{n} 1(y_i = 1) + 1, \ldots, \frac{1}{n} \sum_{i=1}^{n} 1(y_i = C) + 1 \right),$$

and impute x_i^{mis} from $\pi(x_i^{mis} \,|\, y_i, \eta_i)$. Then, compute ζ_j in (2) for $j = 1, \ldots, p$.

4 Results

Using the MCMC algorithm described in the previous section, we fit our model to the PHMRC VA data. We are particularly interested in the improvement that comes from explicitly accounting for dependence between symptoms. We evaluate whether incorporating dependence between symptoms improves prediction of the distribution of deaths by cause in a target population. As a measure of difference based on L_1 distance, we utilize cause specific mortality fraction (CSMF) accuracy, which takes a value in tne [0,1] interval and is

defined as

$$\text{CSMF} = 1 - \frac{\sum_{c=1}^{C} |\pi_0(y = c) - \pi(y = c)|}{2\{1 - \min_{1 \le c \le C} \pi_0(y = c)\}}.$$

An additional consideration in our evaluation is the ability of the method to generalize when the cause of death distribution (and possibly the relationship between symptoms and causes) varies between the training and testing set. This consideration is fundament in the VA setting since obtaining training data is extremely costly. Therefore in many settings, we have limited training data from one geographic area, or site, and then apply the method to predict the distribution of deaths by cause at another site. In evaluating the method, we first consider a scenario where we have target and training data from two different sites, then another survey is conducted in the target site and the new information is added to the training data. To simulate the more realistic generalization test, we divide the PHMRC data into training, target and additional training sets.

First we assume a case where the target and training sites have the same distribution of causes. 1,700 persons (around 20% of the total observations) are randomly selected as data in a target site and the rest are classified as training data. Then the former data are separated into target and additional training sets. We consider three situations where the percentage of additional training data is 0%, 5% or 10% of all the training data, while the target data are fixed. We set $R = 200$ in (3) and generated 5,000 MCMC samples after the initial 500 samples were discarded as a burn-in period, and every 10th sample was saved. We observed that the sample paths were stable, and the sample autocorrelations dropped smoothly. Illustrative examples of the sample plot and the autocorrelation are in the supplementary materials. The top in Figure 5 shows CSMF accuracies for the proposed model with numbers of factors $K = 1, \ldots, 10$ and the conditionally independent Bayesian model. Because the training and target data have the same distribution of causes, we observe little difference even if the additional training information is available. Using the proposed model, as K increases the CSMF accuracy also increases along a convex path to a maximum of about 0.9, a level far above that of the conditional independent model.

Then we consider more realistic cases where the target and training sites have different distributions of causes. We study six cases in which each of the PHMRC sites (AP, Bohol,

Dar, Mexico, Pemba, UP) is treated as a target site and the rest together as training data. Figures 4 show empirical distributions of causes in the target and training sites with L_1 distances. We observe discrepancies between the distributions of deaths by cause for each pair of training and target sites. The L_1 distances for AP and Mexico are relatively small, and Pemba shows the largest gap. We assume no additional training data for Pemba because it has a relatively small number of deaths ($n = 297$).

Figures 5-7 report the boxplots of CSMF accuracies for each site. Unlike the simple case estimates of the CSMFs improve as additional data are added from the target site. This suggests that the estimation of population distributions of deaths by cause can be improved in practice by collecting data in a target site, even if training data are already available from other areas. In all cases the CSMF accuracies obtained by the proposed model are as high or higher than the conditional independent model. For Pemba both the proposed method and the conditional independent model show relatively small CSMF accuracies, probably because there is a large discrepancy between the distributions of deaths by cause for the target and training sites. Still, the proposed model works better. The bottom in Figure 7 indicates averaged CSMF accuracy over all sites except Pemba. The proposed model produces higher values, and inspite of the fact that the conditional independent model improves as more information is added from the target sites, the gap between the two methods actually becomes progressively larger, indicating that the proposed method improves faster. As a function of K, the proposed model consistently produces a concave shape with peaks around $K = 5$.

Turning now to results for our measure of conditional dependence, we estimate conditional mutual information of each item using the proposed model with $K = 5$. For this purpose missing values need to be imputed, but some items show high missing rates as in Figure 2. To obtain robust results we include only items with a missing rate less than 5%. Figure 8 displays boxplots of the estimated conditional mutual information. We observe that several items related to medical history show high conditional dependence, such as *6. Did decedent have AIDS?*, *8. Did decedent have cancer?*, *12. Did decedent have diabetes?*, and *14. Did decedent have heart disease?* Other factors also indicate relatively strong conditional associations, for example *131. Did decedent have any swelling or lump in the breast?*,

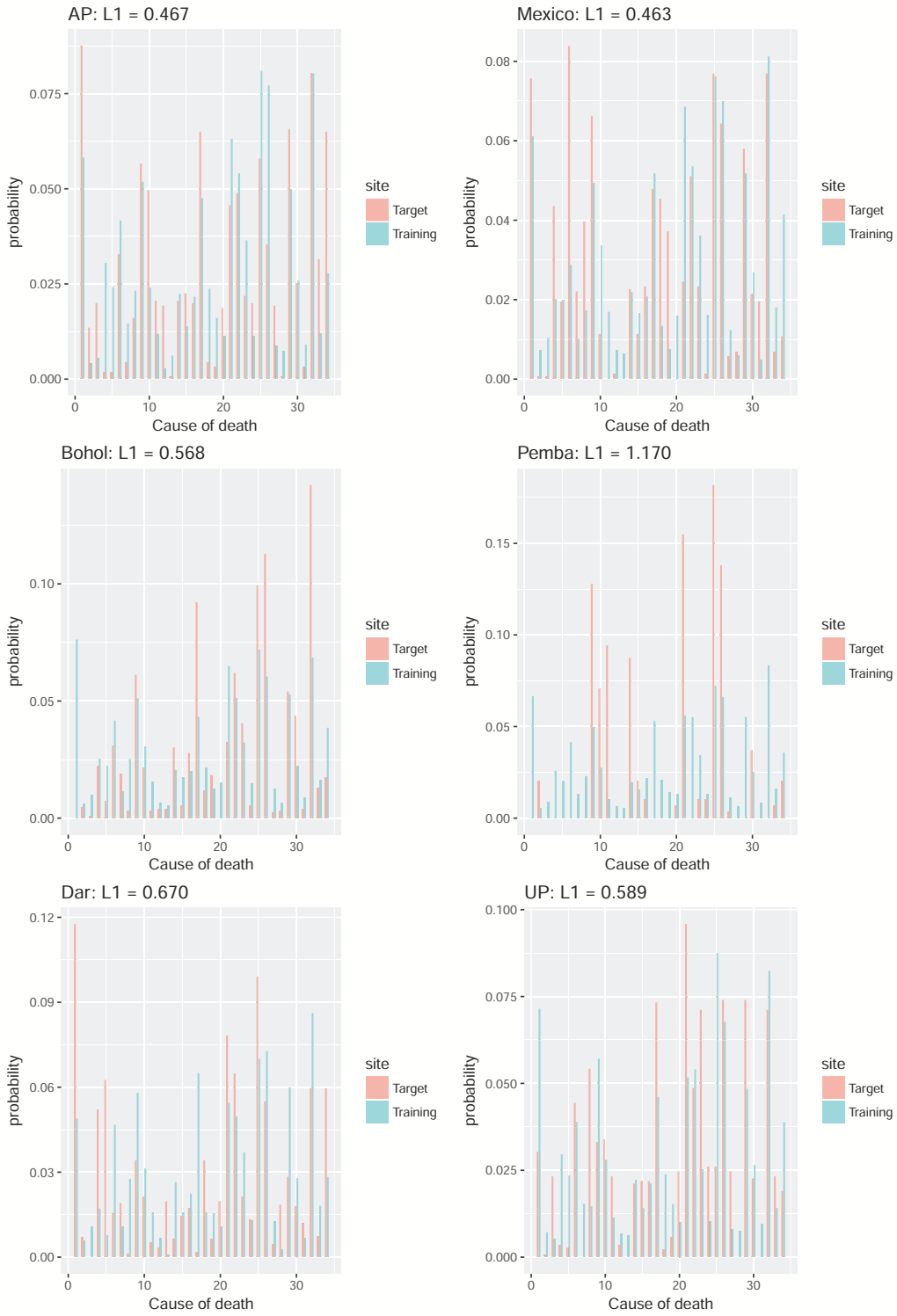

Figure 4: Empirical distirubitons of causes in target and training sites in simulation studies. The target sites are AP, Bohol, Dar (left) and Mexico, Pemba, UP (right). L1 indicates the L_1 distance between the two distributions.

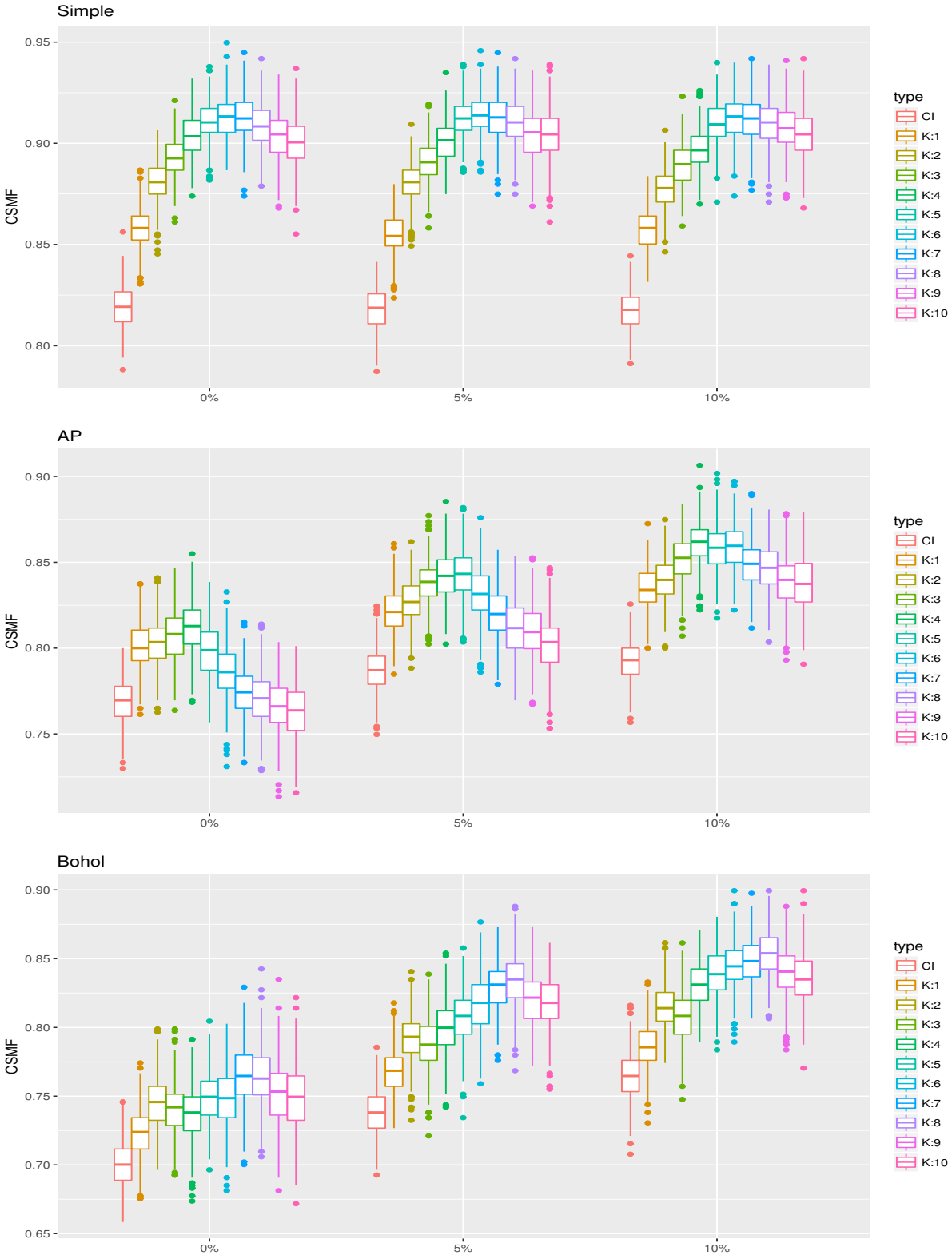

Figure 5: Boxplots of CSMF accuracies for the simple case (above), AP (middle), Bohol (bottom). x-axis shows the percentage of training data from the target site. CI and K means the conditional independent model and the proposed model with K factors.

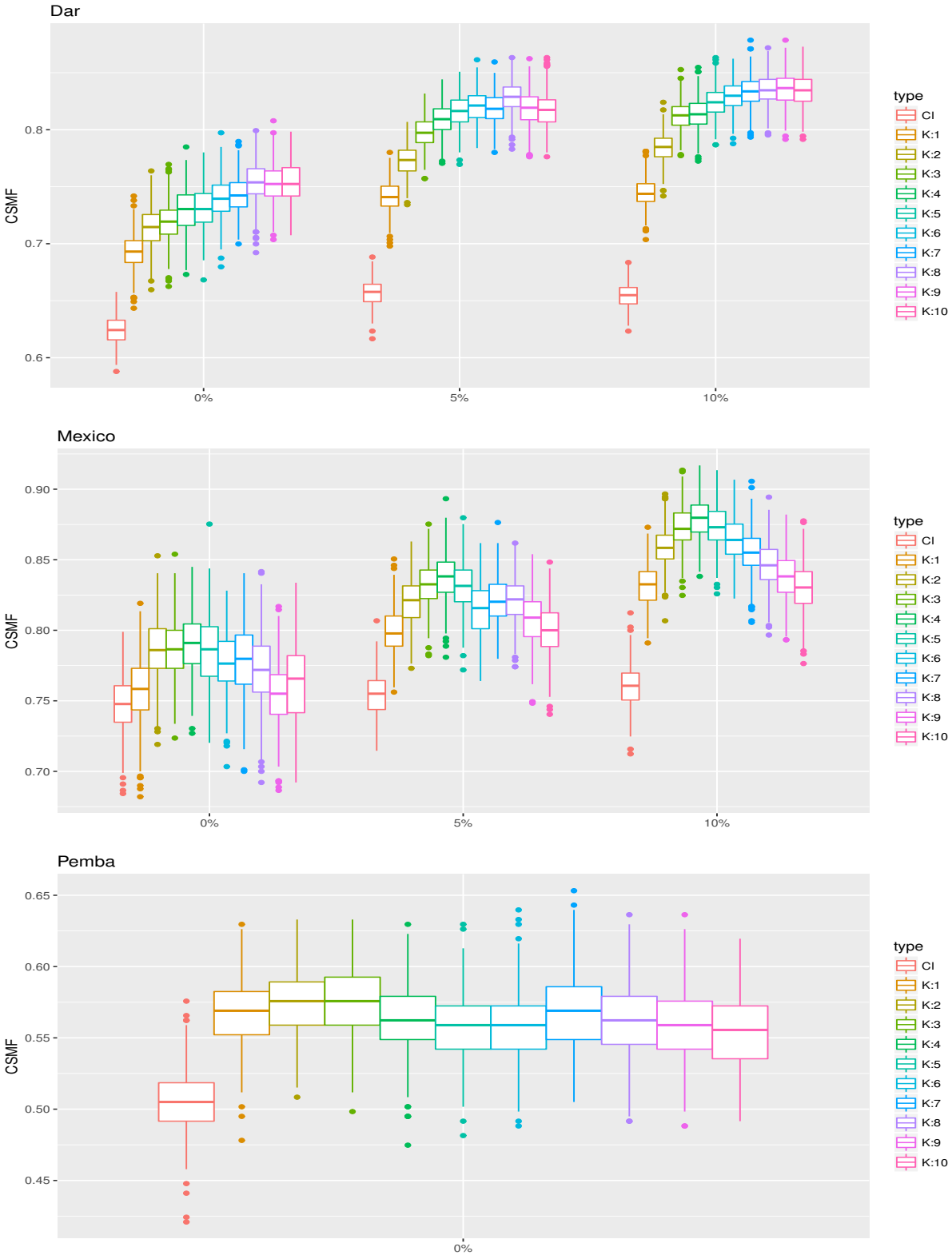

Figure 6: Boxplots of CSMF accuracies for Dar (above), Mexico (middle), Pemba (bottom). x-axis shows the percentage of training data from the target site. CI and K means the conditional independent model and the proposed model with K factors.

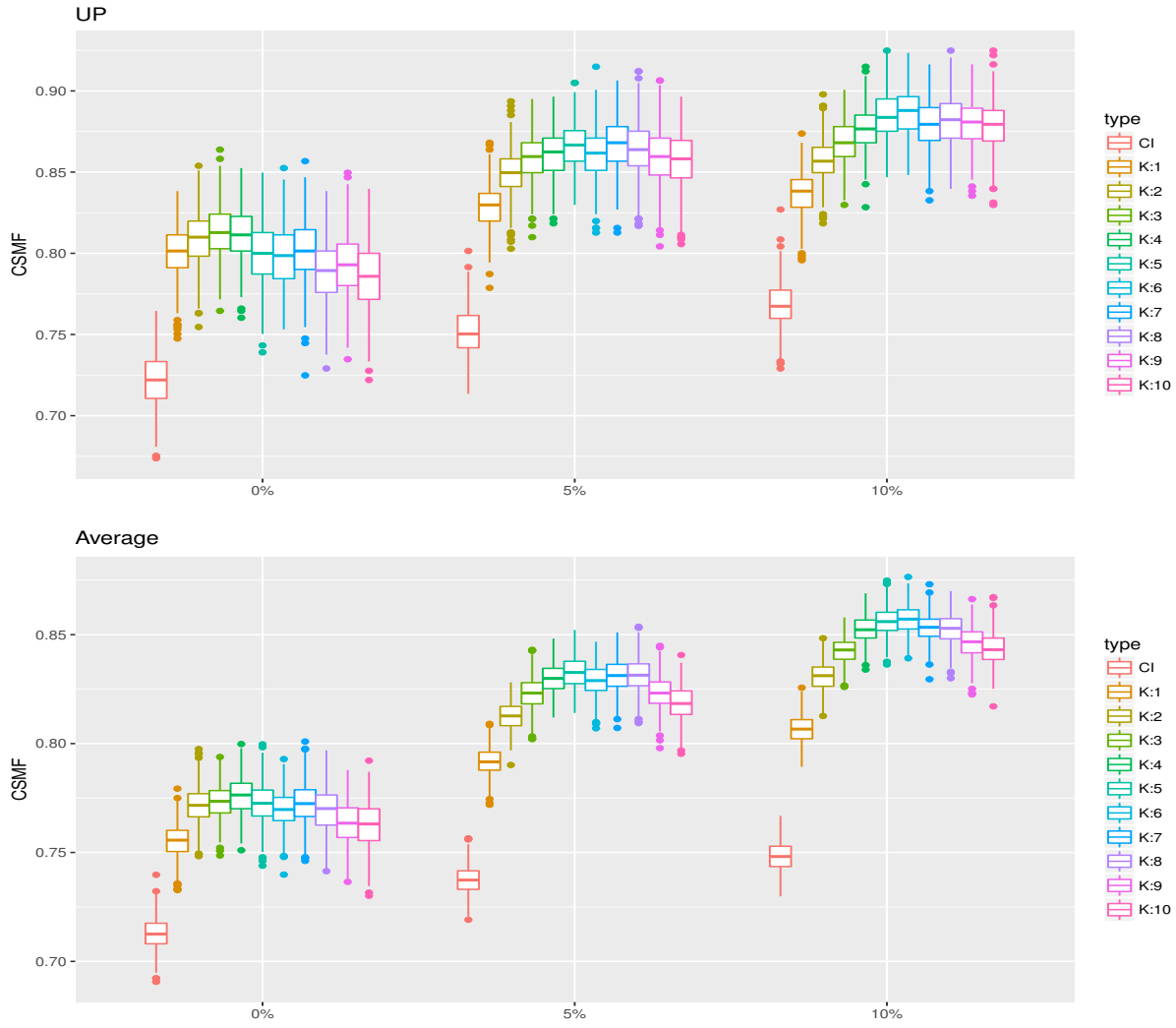

Figure 7: Boxplots of CSMF accuracies for UP (above) and averaged values over all sites except Pemba (bottom). x-axis shows the percentage of training data from the target site. CI and K means the conditional independent model and the proposed model with K factors.

146. Did decedent die within 6 weeks of childbirth?, *159.Did decedent drink alcohol?*, and *171. Was the injury or accident self-inflicted?* Some items mainly report negative values probably because they highly correlate with other questions while they add little new information about causes, leading to additional model complexity and hence degradation of the overall fit. For example predictors 55-57, 108-111 and 149-158 are related to cough, loss of consciousness, and smoking, respectively. These effectively represent the same information, and hence each is redundant given the other questions.

5 Conclusion

We develop a new Bayesian method to estimate distributions of deaths by cause using VA data. The new approach flexibly captures complex interactions among questionnaire items avoiding the restrictive assumption of conditional independence. In the proposed framework, the strength of conditional dependence of symptoms with the causes can be measured as conditional mutual information.

One future direction is to incorporate spatial information into the proposed model. Factors affecting cause of death vary through space depending on geographic characteristics, so two sites that are close to each other should largely share the same factors affecting cause of death. Therefore it may be more efficient to estimate distributions of deaths by cause by weighting more on neighboring areas. In addition the relationship between causes and questionnaire items may depend on space. Although this article assumes the conditional distribution of the symptoms given a cause is constant over space, one can extend it to $\pi(x \mid y, s)$ with spatial information s.

Another direction for future work is to generalize the proposed framework for survey weights. To save cost and time many social surveys employ special data-collection designs such as stratified sampling that produce a biased sample. To adjust for a gap between the sample and the population, survey weights are constructed and distributed along with the data. When faced with data like that, it is necessary to incorporate the weights into statistical models for prediction.

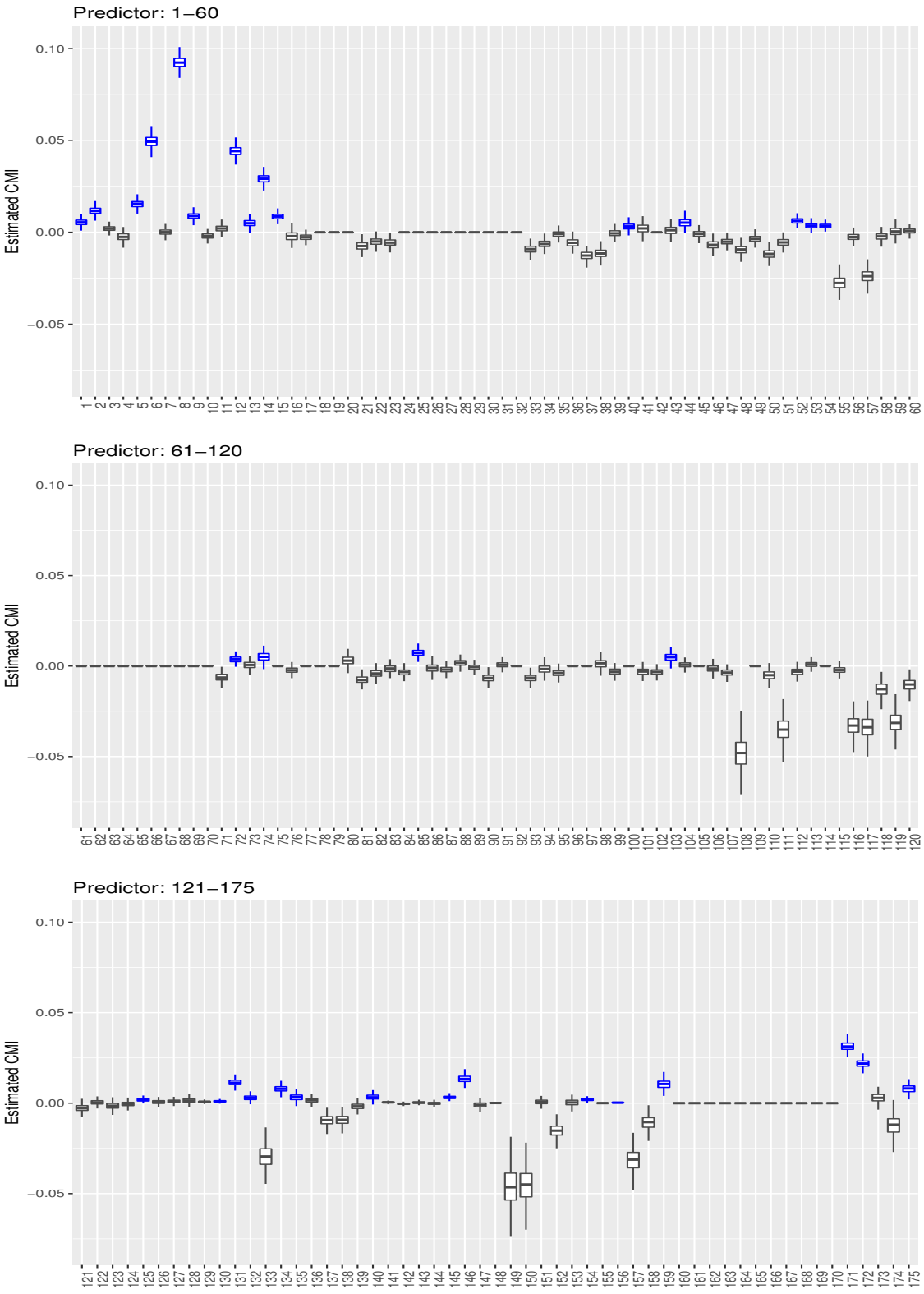

Figure 8: Boxplots of estimated conditional mutual information for symptoms 1-60 (above), 61-120 (middle), 120-175 (bottom). Blue color corresponds to $P[\zeta_j > 0] > 0.95$.

Acknowledgment

This work was supported by The University of Washington eScience Moore/Sloan & WRF Data Science Fellowship, JSPS Grant-in-Aid for Research Activity Start-up 16H06853, and grants K01HD078452 and 1R01HD086227 from the National Institute of Child Health and Human Development (NICHD). The results are generated mainly using Ox (Doornik, 2007). Replication code for the proposed method and the conditional independent model is available on GitHub (https://github.com/kunihama/VA-code.git) .

References

AbouZahr, C., J. Cleland, F. Coullare, S. B. Macfarlane, F. C. Notzon, P. Setel, S. Szreter, R. N. Anderson, A. A. Bawah, A. P. Betrán, F. Binka, K. Bundhamcharoen, R. Castro, T. Evans, X. Figueroa, C. K. George, L. Gollogly, R. Gonzalez, D. R. Grzebien, K. Hill, Z. Huang, T. H. Hull, M. Inoue, R. Jakob, P. Jha, Y. Jiang, R. Laurenti, X. Li, D. Lievesley, A. D. Lopez, D. M. Fat, M. Merialdi, L. Mikkelsen, J. K. Nien, C. Rao, K. Rao, O. Sankoh, K. Shibuya, N. Soleman, S. Stout, V. Tangcharoensathien, P. J. van der Maas, F. Wu, G. Yang, and S. Zhang (2007). The way forward. *Lancet 370*, 1791–1799.

Baiden, F., A. Bawah, S. Biai, F. Binka, T. Boerma, P. Byass, D. Chandramohan, S. Chatterji, C. Engmann, D. Greet, R. Jakob, K. Kahn, O. Kunii, A. D. Lopez, C. J. Murray, B. Nahlen, C. Rao, O. Sankoh, P. W. Setel, K. Shibuya, N. Soleman, L. Wright, and G. Yang (2007). Setting international standards for verbal autopsy. *Bulletin of the World Health Organization 85*, 570–571.

Bloomberg, M. R. and J. Bishop (2015). Understanding death, extending life. *The Lancet 386*(10003), e18–e19.

Boerma, J. T. and S. K. Stansfield (2007). Health statistics now: are we making the right investments? *Lancet 369*, 779–786.

Byass, P., D. Chandramohan, S. J. Clark, L. D'Ambruoso, E. Fottrell, W. J. Graham, A. J. Herbst, A. Hodgson, S. Hounton, K. Kahn, A. Krishnan, J. Leitao, F. Odhiambo, O. A.

Sankoh, and S. M. Tollman (2012). Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool. *Global Health Action 5*, 19281.

Cover, T. M. and J. A. Thomas (2006). *Elements of information theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

de Savigny, D., I. Riley, D. Chandramohan, F. Odhiambo, E. Nichols, S. Notzon, C. AbouZahr, R. Mitra, D. Cobos Muñoz, S. Firth, et al. (2017). Integrating community-based verbal autopsy into civil registration and vital statistics (crvs): system-level considerations. *Global health action 10*(1), 1272882.

Doornik, J. A. (2007). *Object-oriented matrix programming using Ox, 3rd ed.* London: Timberlake Consultants Press and Oxford: www.doornik.com.

Hill, K., A. D. Lopez, K. Shibuya, P. Jha, C. AbouZahr, R. N. Anderson, A. A. Bawah, A. P. Betrán, F. Binka, K. Bundhamcharoen, R. Castro, J. Cleland, F. Coullare, T. Evans, X. C. Figueroa, C. Korah, L. Gollogly, R. Gonzalez, D. R. Grzebien, Z. Huang, T. H. Hull, M. Inoue, R. Jakob, Y. Jiang, R. Laurenti, X. Li, D. Lievesley, D. M. Fat, S. Macfarlane, P. Mahapatra, M. Merialdi, L. Mikkelsen, J. K. Nien, F. C. Notzon, C. Rao, K. Rao, O. Sankoh, P. W. Setel, N. Soleman, S. Stout, S. Szreter, V. Tangcharoensathien, P. J. van der Maas, F. Wu, G. Yang, S. Zhang, and M. Zhou (2007). Interim measures for meeting needs for health sector data: births, deaths, and causes of death. *Lancet 370*, 1726–1735.

Horton, R. (2007). Counting for health. *Lancet 370*, 1526.

James, S. L., A. D. Flaxman, and C. J. Murray (2011). Performance of the Tariff Method: validation of a simple additive algorithm for analysis of verbal autopsies. *Population Health Metrics 9*, 31.

Jha, P. (2014). Reliable direct measurement of causes of death in low-and middle-income countries. *BMC medicine 12*(1), 19.

Joe, H. (1989). Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association 84*, 157–164.

King, G. and Y. Lu (2008). Verbal autopsy methods with multiple causes of death. *Statistical Science 23*, 78–91.

King, G., Y. Lu, and K. Shibuya (2010). Designing verbal autopsy studies. *Population Health Metrics 8*, 19.

Kunihama, T. and D. B. Dunson (2016). Nonparametric Bayes inference on conditional independence. *Biometrika 103*, 35–47.

Lopez, A. D. (1998). Counting the dead in China. *BMJ 317*, 1399–1400.

Lozano, R., A. D. Lopez, C. Atkinson, M. Naghavi, A. D. Flaxman, and C. J. Murray (2011). Performance of physician-certified verbal autopsies: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics 9*, 32.

Mahapatra, P., K. Shibuya, A. D. Lopez, F. Coullare, F. C. Notzon, C. Rao, S. Szreter, C. AbouZahr, R. N. Anderson, A. A. Bawah, A. P. Betrán, F. Binka, K. Bundhamcharoen, R. Castro, J. Cleland, T. Evans, X. C. Figueroa, C. Korah, L. Gollogly, R. Gonzalez, D. R. Grzebien, K. Hill, Z. Huang, T. H. Hull, M. Inoue, R. Jakob, J. P., Y. Jiang, R. Laurenti, X. Li, D. Lievesley, D. M. Fat, S. Macfarlane, M. Merialdi, L. Mikkelsen, J. K. Nien, K. Rao, O. Sankoh, P. W. Setel, N. Soleman, S. Stout, V. Tangcharoensathien, P. J. van der Maas, F. Wu, G. Yang, S. Zhang, and M. Zhou (2007). Civil registration systems and vital statistics: successes and missed opportunities. *Lancet 370*, 1653–63.

Maher, D., S. Biraro, V. Hosegood, R. Isingo, T. Lutalo, P. Mushati, B. Ngwira, M. Nyirenda, J. Todd, and B. Zaba (2010). Translating global health research aims into action: the example of the ALPHA network. *Tropical Medicine & International Health 15*, 321–328.

Mathers, C. D., D. M. Fat, M. Inoue, C. Rao, and A. D. Lopez (2005). Counting the dead and what they died from: an assessment of the global status of cause of death data. *Bulletin of the World Health Organization 83*, 171–177.

McCormick, T. H., Z. Li, C. Calvert, A. C. Crampin, K. Kahn, and S. J. Clark (2016). Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association 111*, 1036–1049.

Miasnikof, P., V. Giannakeas, M. Gomes, L. Aleksandrowicz, A. Y. Shestopaloff, D. Alam, S. Tollman, A. Samarikhalaj, and P. Jha (2015). Naive Bayes classifiers for verbal autopsies: Comparison to physician-based classification for 21,000 child and adult deaths. *BMC Medicine 13*, 286.

Mikkelsen, L., D. E. Phillips, C. AbouZahr, P. W. Setel, D. De Savigny, R. Lozano, and A. D. Lopez (2015). A global assessment of civil registration and vital statistics systems: monitoring data quality and progress. *The Lancet 386*(10001), 1395–1406.

Murray, C. J., S. L. James, J. K. Birnbaum, M. K. Freeman, R. Lozano, and A. D. Lopez (2011). Simplified Symptom Pattern Method for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics 9*, 30.

Murray, C. J., A. D. Lopez, R. Black, R. Ahuja, S. M. Ali, A. Baqui, L. Dandona, E. Dantzer, V. Das, U. Dhingra, A. Dutta, W. Fawzi, A. D. Flaxman, S. Gómez, B. Hernández, R. Joshi, H. Kalter, A. Kumar, V. Kumar, R. Lozano, M. Lucero, S. Mehta, B. Neal, S. L. Ohno, R. Prasad, D. Praveen, Z. Premji, D. Ramírez-Villalobos, H. Remolador, I. Riley, M. Romero, M. Said, D. Sanvictores, S. Sazawal, and V. Tallo (2011). Population Health Metrics Research Consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Population Health Metrics 9*, 27.

Murray, C. J., A. D. Lopez, D. M. Feehan, S. T. Peter, and G. Yang (2007). Validation of the Symptom Pattern Method for analyzing verbal autopsy data. *PLoS Medicine 4*, e327.

Navarro, D. (2015). *Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.5).* Adelaide, Australia: University of Adelaide. http://ua.edu.au/ccs/teaching/lsr.

Nichols, E. K., P. Byass, D. Chandramohan, S. J. Clark, A. D. Flaxman, R. Jakob, J. Leitao, N. Maire, C. Rao, I. Riley, et al. (2018). The WHO 2016 verbal autopsy instrument: An international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0. *PLoS Medicine 15*(1), e1002486.

Phillips, D., R. Lozano, M. Naghavi, C. Atkinson, D. Gonzalez-Medina, L. Mikkelsen, C. Murray, and A. Lopez (2014). A composite metric for assessing data on mortality and causes of death: the vital statistics performance index. *Population Health Metrics 12*, 14.

Phillips, D. E., C. AbouZahr, A. D. Lopez, L. Mikkelsen, D. De Savigny, R. Lozano, J. Wilmoth, and P. W. Setel (2015). Are well functioning civil registration and vital statistics systems associated with better health outcomes? *The Lancet 386*(10001), 1386–1394.

PHMRC (2013). Population Health Metrics Research Consortium gold standard verbal autopsy data 2005-2011. http://ghdx.healthdata.org/record/population-health-metrics-research-consortium-gold-standard-verbal-autopsy-data-2005-2011.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, https://www.R-project.org/.

Ruzicka, L. T. and A. D. Lopez (1990). The use of cause-of-death statistics for health situation assessment: national and international experiences. *World Health Statistics Quarterly 43*, 249–258.

Sankoh, O. and P. Byass (2012). The INDEPTH Network: filling vital gaps in global epidemiology. *International Journal of Epidemiology 41*, 579–588.

Serina, P., I. Riley, A. Stewart, S. L. James, A. D. Flaxman, R. Lozano, B. Hernandez, M. D. Mooney, R. Luning, R. Black, R. Ahuja, N. Alam, S. S. Alam, S. M. Ali, C. Atkinson, A. H. Baqui, H. R. Chowdhury, L. Dandona, R. Dandona, E. Dantzer, G. L. Darmstadt, V. Das, U. Dhingra, A. Dutta, W. Fawzi, M. Freeman, S. Gomez, H. N. Gouda, R. Joshi, H. D. Kalter, A. Kumar, V. Kumar, M. Lucero, S. Maraga, S. Mehta, B. Neal, S. L. Ohno, D. Phillips, K. Pierce, R. Prasad, D. Praveen, Z. Premji, D. Ramirez-Villalobos, P. Rarau, H. Remolador, M. Romero, M. Said, D. Sanvictores, S. Sazawal, P. K. Streatfield, V. Tallo, A. Vadhatpour, M. Vano, C. J. L. Murray, and A. D. Lopez (2015). Improving performance of the Tariff Method for assigning causes of death to verbal autopsies. *BMC Medicine 13*, 291.

Setel, P. W., S. B. Macfarlane, S. Szreter, L. Mikkelsen, P. Jha, S. Stout, C. AbouZahr, and M. of Vital Events (2007). A scandal of invisibility: making everyone count by counting everyone. *Lancet 370*, 1569–1577.

Soleman, N., D. Chandramohan, and K. Shibuya (2006). Verbal autopsy: current practices and challenges. *Bulletin of the World Health Organization 84* (3), 239–245.

World Health Organization (2012, accessed 2018-02). *Verbal Autopsy Standards: The 2012 WHO verbal autopsy instrument*. https://goo.gl/bQXXhG.

World Health Organization (2017, accessed 2018-02). *Verbal Autopsy Standards: The 2016 WHO verbal autopsy instrument*. https://goo.gl/Hgt6es.

Wyner, A. D. (1978). A definition of conditional mutual information for arbitrary ensembles. *Information and Control 38*, 51–59.

Yang, G., J. Hu, K. Q. Rao, J. Ma, C. Rao, and A. D. Lopez (2005). Mortality registration and surveillance in China: History, current situation and challenges. *Population Health Metrics 3*, 3.

Supplementary materials for "Bayesian factor models for probabilistic cause of death assessment with verbal autopsies"

1 MCMC convergence

To investigate posterior convergence, we show illustrative examples of the sample paths and autocorrelations of the MCMC sample in Section 4. Figures 1-4 report results by the conditional independent model and the proposed model with $K = 2, 4, 6, 8, 10$ for the senario without additional training data.

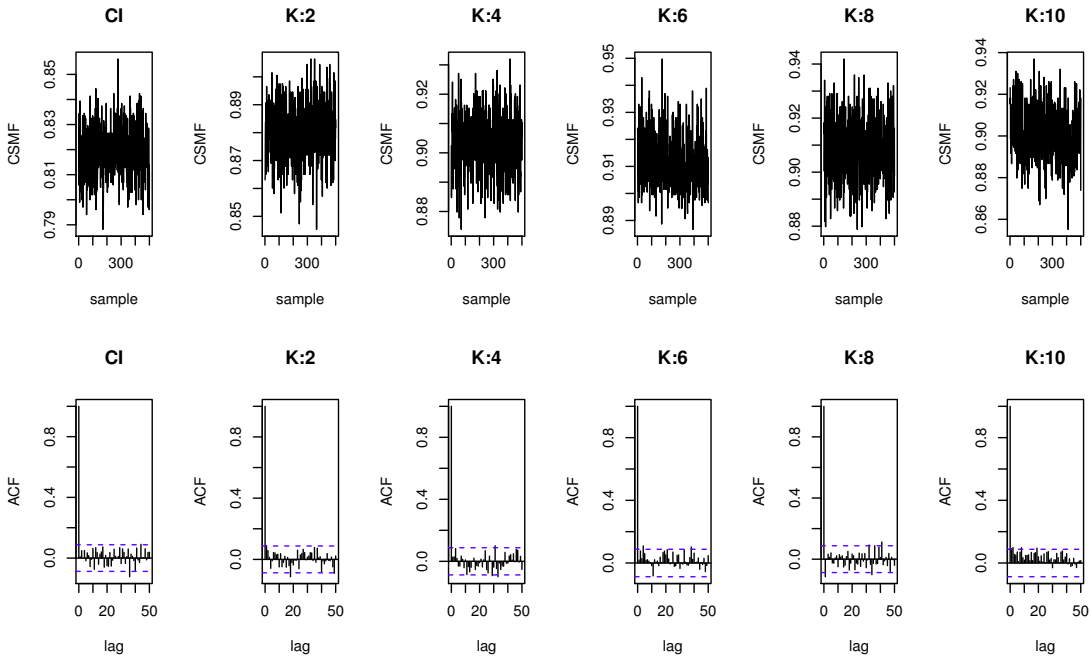

Figure 1: Sample paths and autocorrelations for the simple case. CI, K:2, K:4, K:6, K:8, K:10 represent the conditional independent model and the proposed model with $K = 2, 4, 6, 8, 10$.

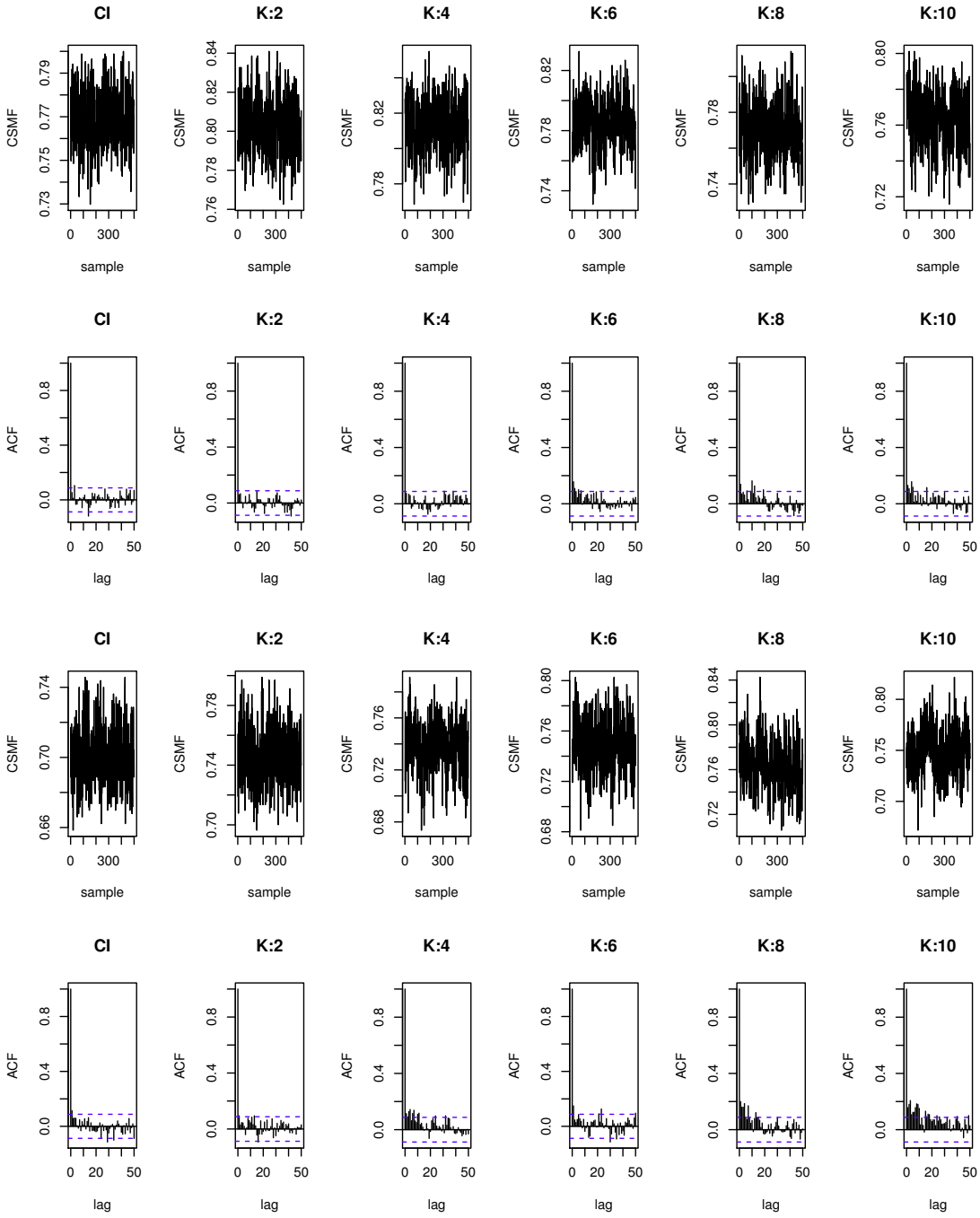

Figure 2: Sample paths and autocorrelations for AP (upper half) and Bohol (lower half). CI, K:2, K:4, K:6, K:8, K:10 represent the conditional independent model and the proposed model with $K = 2, 4, 6, 8, 10$.

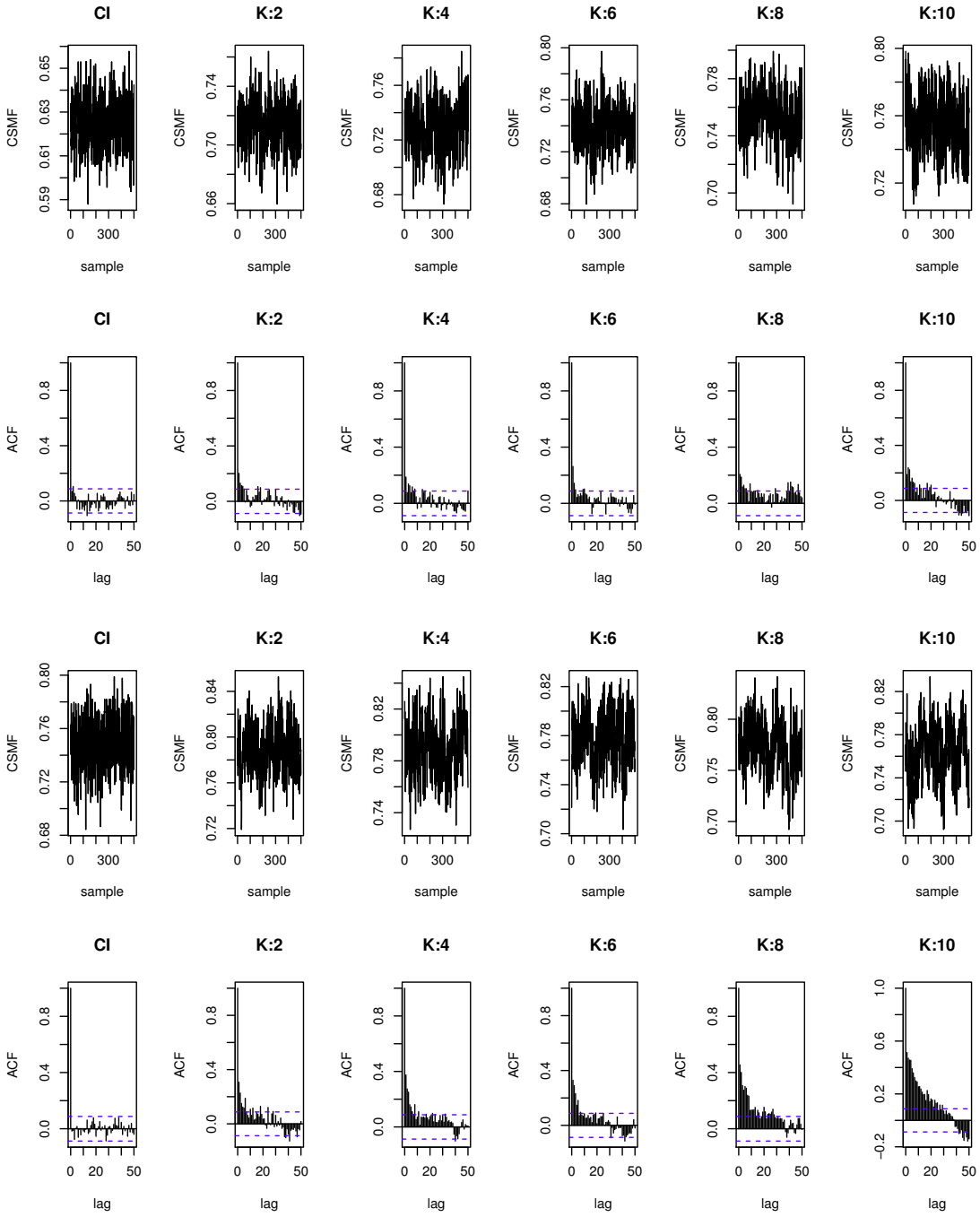

Figure 3: Sample paths and autocorrelations for Dar (upper half) and Mexico (lower half). CI, K:2, K:4, K:6, K:8, K:10 represent the conditional independent model and the proposed model with $K = 2, 4, 6, 8, 10$.

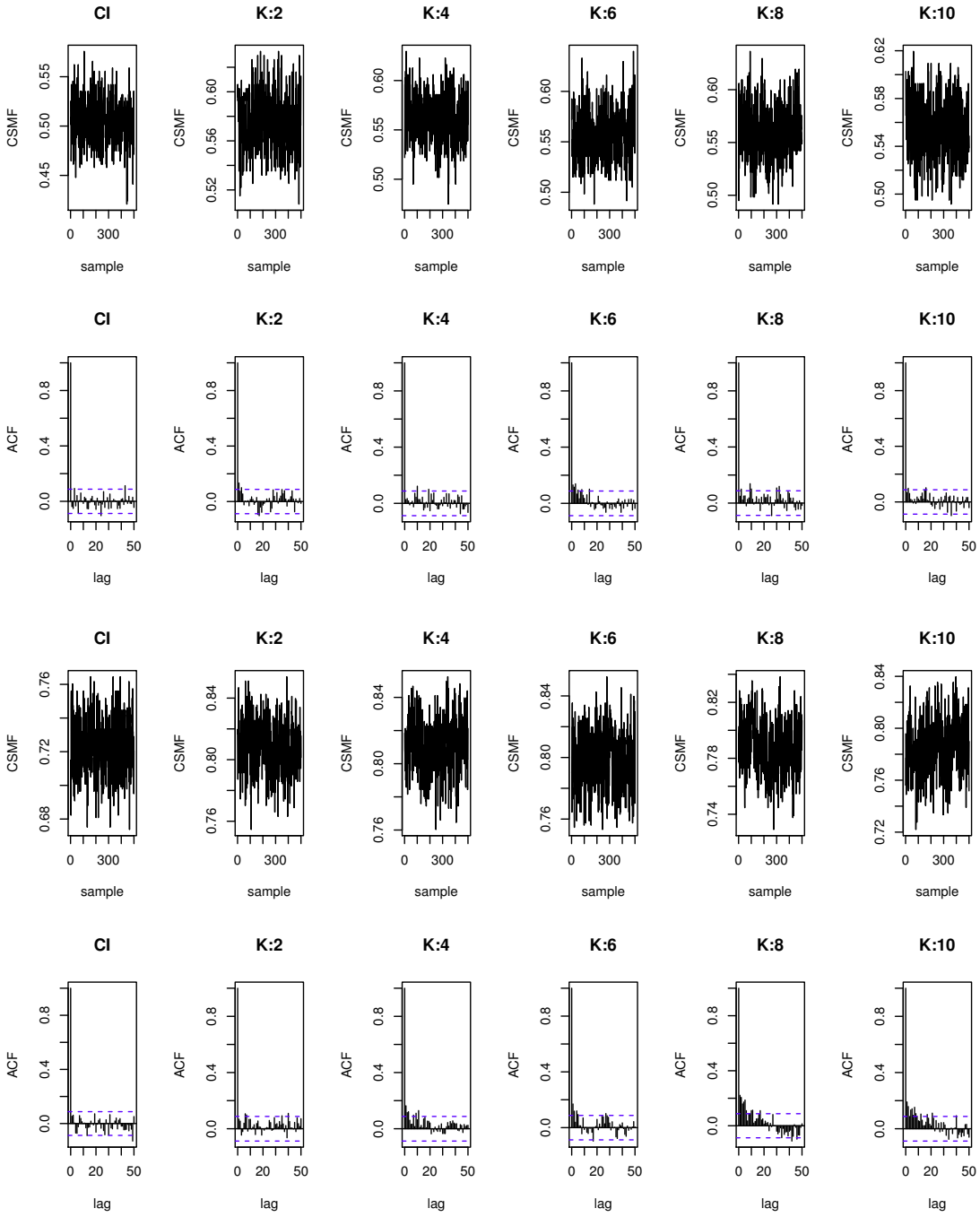

Figure 4: Sample paths and autocorrelations for Pemba (upper half) and UP (lower half). CI, K:2, K:4, K:6, K:8, K:10 represent the conditional independent model and the proposed model with $K = 2, 4, 6, 8, 10$.