Supplementary Material to “Using Bayesian Latent Gaussian Graphical Models to Infer Symptom Associations in Verbal Autopsies”

Zehang Richard Li[*], Tyler H. McComick[†] and Samuel J. Clark[‡]

1 Derivation of the spike-and-slab prior

The proposed prior distribution for $\boldsymbol{R}$ can be factored into two parts,

$$\begin{aligned} p(\boldsymbol{R}|\boldsymbol{\delta}) &= C_{\boldsymbol{\delta}}^{-1}|\boldsymbol{R}|^{-(p+1)}\prod_{j<k}\text{Normal}(r^{jk}|0, v^2_{\delta_{jk}})\prod_{j}\text{Exp}(r^{jj}|\lambda/2)\mathbf{1}_{\boldsymbol{R}\in R^+} \\ p(\boldsymbol{\delta}|\pi_\delta) &\propto C_{\boldsymbol{\delta}}\prod_{j<k}\pi_\delta^{\delta_{jk}}(1-\pi_\delta)^{1-\delta_{jk}} \end{aligned}$$

where $C_{\boldsymbol{\delta}}$ is a normalizing constant. First we show that $C_{\boldsymbol{\delta}} < \infty$ so that the prior distribution is proper. We note

$$\begin{aligned} C_{\boldsymbol{\delta}} &= C\int_{R^+}|\boldsymbol{R}|^{-(p+1)}\prod_{j<k}\exp(-(r^{jk})^2/2v^2_{\delta_{jk}})\prod_{j}\exp(-\lambda r^{jj}/2)d\boldsymbol{R} \\ &\leq C\int_{R^+}|\boldsymbol{R}|^{-(p+1)}\prod_{j}\exp(-\lambda r^{jj}/2)d\boldsymbol{R} \\ &= C\int_{R^+}|\boldsymbol{R}|^{-(p+1)}\prod_{j}(r^{jj})^{-\frac{p+1}{2}}\prod_{j}\exp(-\lambda r^{jj}/2+\frac{p+1}{2}\log(r^{jj}))d\boldsymbol{R} \end{aligned}$$

Since $\exp(-\lambda r^{jj}/2 + \frac{p+1}{2}\log(r^{jj}))$ is a non-negative function of r^{jj}, and has a global maximum at $r^{jj} = (p+1)/\lambda$, and C is a positive constant, we have

$$C_{\boldsymbol{\delta}} \leq C'\int_{R^+}|\boldsymbol{R}|^{-(p+1)}\prod_{j}(r^{jj})^{-\frac{p+1}{2}}d\boldsymbol{R},$$

where the constant $C' < \infty$, and $\int_{R^+}|\boldsymbol{R}|^{-(p+1)}\prod_j(r^{jj})^{-\frac{p+1}{2}}d\boldsymbol{R} < \infty$ as well since it is proportional to the marginally uniform prior of $\boldsymbol{R}$ derived from the Wishart distribution. Therefore the normalizing constant $C_\delta < \infty$, and the prior is proper.

In order to obtain the prior distribution on the expanded precision matrix $\boldsymbol{\Omega} = (\boldsymbol{DRD})^{-1}$, we put prior on the marginal expansion parameter $\boldsymbol{D}$ with a prior distribution so that $p(d_j^2|\boldsymbol{R})$ is an inverse Gamma distribution with shape and rate parameter being $((p+1)/2, 1/2)$, we have

$$p(\boldsymbol{D}|\boldsymbol{R}) \propto \prod_{j} d_j^{-(p+2)}\exp(\frac{1}{2d_j^2})$$

By definition, we know $r^{jk} = \omega_{jk}d_jd_k$. The Jacobian of the transformation from $\boldsymbol{\Omega}$ to $\boldsymbol{\Sigma}$ is $|\boldsymbol{\Sigma}|^{-p-1}$. Under the transformation from $\boldsymbol{\Sigma}$ to $(\boldsymbol{D},\boldsymbol{R})$, the Jacobian is given by $2^p\prod d_j^p$. Putting them together, we can derive

$$p(\boldsymbol{\Omega}|\boldsymbol{\delta}) = p(\boldsymbol{R}|\delta)p(\boldsymbol{D}|\boldsymbol{R})|\mathcal{J}|$$

[*] Department of Biostatistics, Yale School of Public Health, New Haven, CT, zehang.li@yale.edu

[†] Department of Statistics and Department of Sociology, University of Washington, Seattle, WA, tylermc@uw.edu

[‡] Department of Sociology, The Ohio State University, Columbus, OH, work@samclark.net

$$
\begin{aligned}
&\propto && C_{\boldsymbol{\delta}}^{-1}|\boldsymbol{R}|^{-(p+1)}\prod_{j<k}\exp(-(r^{jk})^2/2v^2_{\delta_{jk}})\prod_{j}\exp(-\lambda r^{jj}/2)\prod_{j}d_j^{-(p+2)}\exp(\frac{1}{2d_j^2})|\boldsymbol{R}|^{p+1}\prod d_j^{p+2}\\
&= && C_{\boldsymbol{\delta}}^{-1}\prod_{j<k}\exp(-\frac{\omega_{jk}^2}{2v^2_{\delta_{jk}}/d_j^2d_k^2})\prod_{j}\exp(-\frac{\lambda d_j^2}{2}\omega_{jj})\prod_{j}\exp(\frac{1}{2d_j^2}),
\end{aligned}
$$

where $d_j = \sigma_j$ is the square root of the k-th diagonal element of $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$, i.e.,

$$
p(\boldsymbol{\Omega}|\boldsymbol{\delta}) \propto C_{\boldsymbol{\delta}}^{-1}\prod_{j<k}\exp(-\frac{\omega_{jk}^2}{2v^2_{\delta_{jk}}/\sigma_j^2\sigma_k^2})\prod_{j}\exp(-\frac{\lambda\sigma_j^2}{2}\omega_{jj})\prod_{j}\exp(\frac{1}{2\sigma_j^2})
$$

2 Comparing spike-and-slab with Wishart prior

Since the proposed method is heavily based on the spike-and-slab prior for the precision matrix (Wang, 2015), $\boldsymbol{\Omega}$, we first describe the spike-and-slab prior on the precision matrix, and compare it to other commonly used prior families in this section. Wang (2015) defines the spike-and-slab prior as

$$
\begin{aligned}
p(\boldsymbol{\Omega}|\delta) &\propto C_{\boldsymbol{\delta}}^{-1}\prod_{j<k}\text{Normal}(\omega_{jk}|0, v^2_{\delta_{jk}})\prod_{j}\text{Exp}(\omega_{jj}|\lambda/2)\mathbf{1}_{\Omega\in M^+}\\
p(\delta|\pi_\delta) &\propto C_{\boldsymbol{\delta}}\prod_{j<k}\pi_\delta^{\delta_{jk}}(1-\pi_\delta)^{1-\delta_{jk}}
\end{aligned}
$$

where M^+ denotes the space of positive definite matrices, δ_{jk} are latent indicator variables for each ω_{jk} related to their size (large or small), π_δ is the prior sparsity parameter, and $v_1 \gg v_0$ imposes different levels of shrinkage for the elements drawn from the "slab" and "spike" prior distributions respectively. Conditional on the binary indicators δ_{jk}, this representation shrinks the elements of $\boldsymbol{\Omega}$ differently: a very small v_0 allows us to strongly shrink elements in $\boldsymbol{\Omega}$ to 0 if they are small in scale, and a larger v_1, i.e. a more dispersed prior distribution, shrinks the larger elements only slightly and thus leads to less bias.

Due to the positive definiteness constraint, the normalizing constant for this prior distribution of $\boldsymbol{\Omega}$ is intractable. We glean insights about this prior distribution by simulating from the prior using the MCMC steps described in Wang (2015). Figure 1 shows the induced marginal prior distribution on $\boldsymbol{R}$ and $\boldsymbol{R}^{-1}$ under a complete graph and an $AR(2)$ graph respectively. In the complete graph case when the marginal shrinkage parameter v_1 is large, the marginal prior on $\boldsymbol{R}$ and $\boldsymbol{R}^{-1}$ induced by this spike-and-slab distribution becomes very similar to that of the marginal uniform prior. This is not surprising as it can be seen directly from the marginal distribution on the matrix elements of $\boldsymbol{\Omega}$ as well. For the j-th column of $\boldsymbol{\Omega}$, the spike-and-slab prior induces the conditional prior distribution on $\boldsymbol{\omega}_{[j,-j]}$ and the Schur complement $\omega_{j|-j} = \omega_{jj} - \boldsymbol{\omega}_{[j,-j]}^T\boldsymbol{\Omega}_{[-j,-j]}^{-1}\boldsymbol{\omega}_{[j,-j]}$ to be

$$
\begin{aligned}
\boldsymbol{\omega}_{[j,-j]}|\boldsymbol{\Omega}_{[-j,-j]} &\sim \text{Normal}(\mathbf{0}, (\lambda\boldsymbol{\Omega}_{[-j,-j]}^{-1} + \text{diag}(\boldsymbol{V}_{[j,-j]}^{-1}))^{-1})\\
\omega_{j|-j}|\boldsymbol{\Omega}_{[-j,-j]} &\sim \text{Gamma}\left(1, \frac{\lambda}{2}\right)
\end{aligned}
$$

where $\boldsymbol{V} = \{v^2_{\delta_{jk}}\}_{jk}$ is the matrix of the "penalization" parameters determined by v_0, v_1 and a given graph. This resembles the conditional prior distribution under the Wishart distribution in the previous section, i.e. when $\boldsymbol{\Omega} \sim \text{Wishart}(p+1, \boldsymbol{I}_p)$, the marginal prior distribution for the same quantities are

$$
\begin{aligned}
\boldsymbol{\omega}_{[j,-j]}|\boldsymbol{\Omega}_{[-j,-j]} &\sim \text{Normal}(\mathbf{0}, \boldsymbol{\Omega}_{[-j,-j]})\\
\omega_{j|-j}|\boldsymbol{\Omega}_{[-j,-j]} &\sim \text{Gamma}\left(1, \frac{1}{2}\right)
\end{aligned}
$$

The Wishart prior induced on $\boldsymbol{\omega}_{[j,-j]}$ is the limiting case in the spike-and-slab prior as $v_0 = v_1 \to \infty$ and $\lambda = 1$. The spike-and-slab prior can be viewed, therefore, as a shrinkage prior in the middle ground between

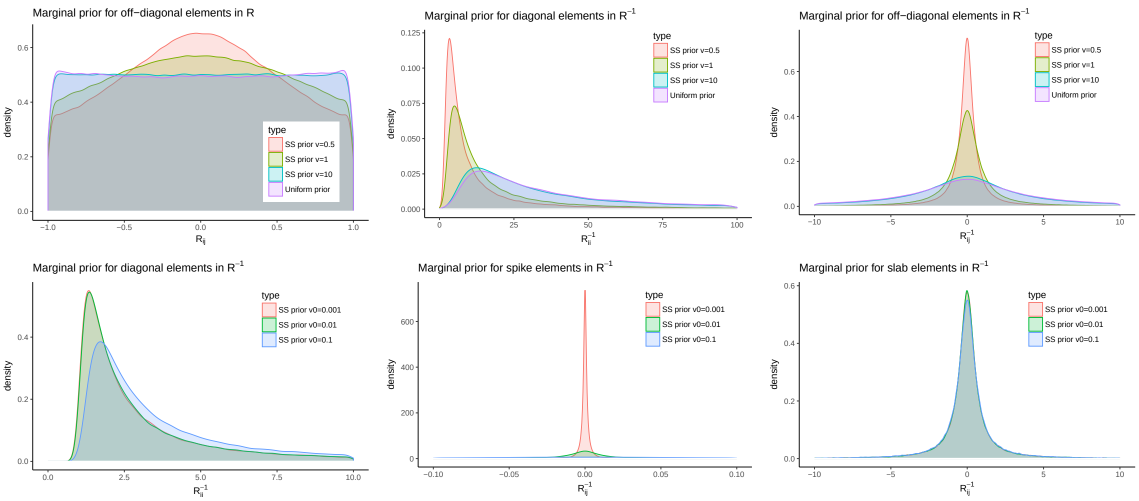


Figure 1: Marginal priors for $\boldsymbol{R}$ and $\boldsymbol{R}^{-1}$. Different marginal priors induced by the spike-and-slab prior on $\boldsymbol{\Omega}$ with $p = 50$ and $\lambda = 2$. **Top row**: marginal priors conditional on a complete graph, i.e. $v_0 = v_1$. Left: off-diagonal elements $\boldsymbol{R}_{ij}, i \neq j$. Middle: diagonal elements $\boldsymbol{R}^{-1}_{ii}$. Right: off-diagonal elements $\boldsymbol{R}^{-1}_{ij}, i \neq j$. **Bottom row**: marginal priors conditional on a fixed $AR(2)$ graph with fixed $v_1 = 1$ and varying v_0 values. Left: diagonal elements $\boldsymbol{R}^{-1}_{ii}$. Middle: Non-zero off-diagonal elements (slab) $\boldsymbol{R}^{-1}_{ij}, i \neq j$. Right: Zero off-diagonal elements (spike) $\boldsymbol{R}^{-1}_{ij}, i \neq j$. The densities are derived from sampling $2,000$ draws using MCMC from the prior distribution after $2,000$ iterations of burn-in.

the Wishart prior and G-Wishart prior where off-diagonal contains exact zeros, while sharing both the easy computational properties of the former and the graph interpretation of the latter.

For the proposed prior on the correlation matrix, we can exam such induced conditional priors in a similar fashion. If we denote $\boldsymbol{\Theta} = \boldsymbol{R}^{-1}$, then $\theta_{j|-j} = 1$, and $\boldsymbol{\theta}_{[j,-j]}|\boldsymbol{\Theta}_{-j,-j]}$ follows similar distribution

$$\boldsymbol{\theta}_{[j,-j]}|\boldsymbol{\Theta}_{[-j,-j]} \quad \sim \quad \text{Normal}(\boldsymbol{0}, (\lambda\boldsymbol{\Theta}^{-1}_{[-j,-j]} + \text{diag}(\boldsymbol{V}^{-1}_{[j,-j]}))^{-1})$$

in the constrained space that $\boldsymbol{\Theta}$ is a inverse correlation matrix. This conditional density also can help guide the choice of the hyperparameters, by comparing λ, v_0, and v_1 to $\boldsymbol{\Theta}^{-1}_{[-j,-j]}$. The scale of $\boldsymbol{\Theta}^{-1}_{[-j,-j]}$ is easy to comprehend, since $\boldsymbol{\Theta}^{-1}_{[-j,-j]} = \boldsymbol{R}_{[-j,-j]} - \boldsymbol{r}^T_{[j,-j]}\boldsymbol{r}_{[j,-j]}$. The linear constraints may render the choice of hyperparameters not straightforward when the edge probability is larger. Nevertheless, we can see from Figure 2 that both the spike-and-slab distributions still changes as expected when we fix all but one parameters, and behaves marginally similar to the spike-and-slab prior for the precision matrix.

3 Implied prior sparsity with different hyperparameters

In this section, we provide more prior simulation results to facilitate the choice of λ, v_0, v_1, and $\pi_{\boldsymbol{\delta}}$. Figure 3 illustrates our approach in understanding how these 4 parameters jointly imply the prior sparsity. It can be seen that small λ and extremely small v_0 usually leads to denser prior graph unless v_1 is also small, which defeats the purpose of using the continuous mixture prior. We choose to use $\lambda = 10$, $v_0 = 0.01$, $v_1/v_0 = 100$, and $\pi_{\boldsymbol{\delta}} = 0.0001$ in our experiments. In general, for the prior edge probability to be calibrated between 0.05 to 0.2, we believe the model is not very sensitive to parameters in the close range to our choices.

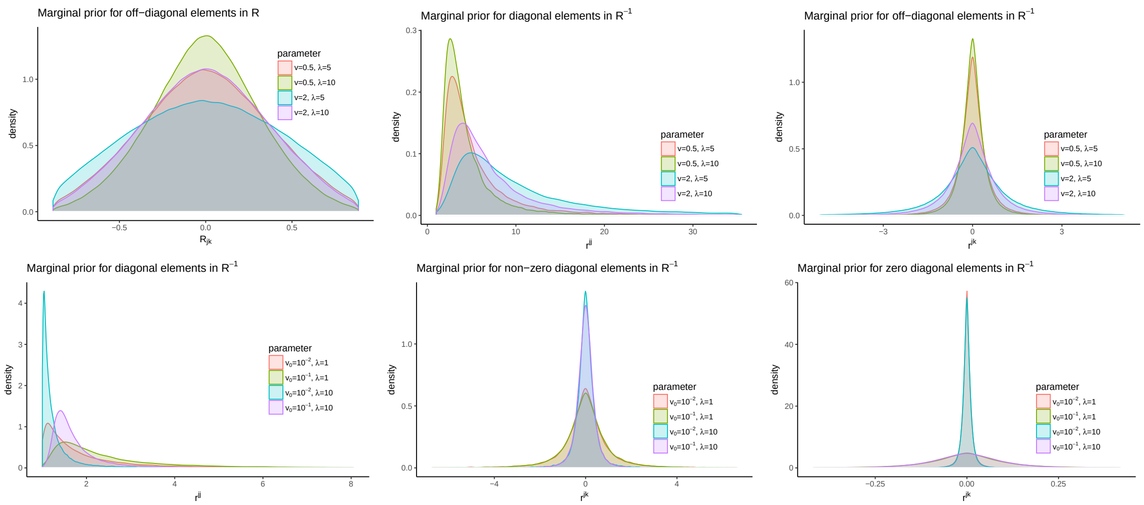


Figure 2: Marginal priors for $\boldsymbol{R}$ and $\boldsymbol{R}^{-1}$. Different marginal priors induced by the spike-and-slab prior on $\boldsymbol{R}$ with $p = 50$. **Top row**: marginal priors conditional on a complete graph, i.e. $v_0 = v_1$. Left: off-diagonal elements $\boldsymbol{R}_{jk}, j \neq k$. Middle: diagonal elements r^{jj}. Right: off-diagonal elements $r^{jk}, j \neq k$. **Bottom row**: marginal priors conditional on a fixed $AR(1)$ graph with fixed $v_1 = 1$ and varying v_0 and λ values. Left: diagonal elements r^{jj}. Middle: Non-zero off-diagonal elements (slab) $r^{jk}, j \neq k$. Right: Zero off-diagonal elements (spike) $r^{jk}, j \neq k$. The densities are derived from sampling $2,000$ draws using MCMC from the prior distribution after $2,000$ iterations of burn-in.

4 Posterior inference for the classification model

This section describes the inference procedure for the model presented in Section 2 of the main paper. The steps are mostly similar to Section 3.2 of the paper.

Update $\boldsymbol{Z}$ and $\boldsymbol{\Lambda}$. This first two steps are the same as in Section 3.2 of the main paper, except replacing $\boldsymbol{\mu}$ to the corresponding $\boldsymbol{\mu}_c$.

Update $\boldsymbol{\mu}$. The conditional posterior distribution for the mean parameters is also multivariate normal,

$$\boldsymbol{\mu}_c | \boldsymbol{Y}, \tilde{\boldsymbol{R}}, \boldsymbol{X} \sim \text{Normal}\left((\frac{1}{\sigma^2}\boldsymbol{I}_p + n_c\tilde{\boldsymbol{R}}^{-1})^{-1}(\frac{1}{\sigma^2}\boldsymbol{\mu}_{0c} + n_c\tilde{\boldsymbol{R}}^{-1}\bar{z}_c), (\frac{1}{\sigma^2}\boldsymbol{I}_p + n_c\tilde{\boldsymbol{R}}^{-1})^{-1} \right)$$

where $n_c = \sum_i \mathbf{1}_{y_i = c}$ and $\bar{z}_c = \sum_{i:y_i=c} \boldsymbol{Z}_i$.

Update $\boldsymbol{R}$. To update the latent correlation matrix, we first draw the working expansion and expand the observations in the same way as Section 3.2 of the main paper. The rescaled sample covariance matrix is $\boldsymbol{S} = \sum_{i=1}^{n}(W_i - \boldsymbol{D}\boldsymbol{\mu}_{y_i})'\boldsymbol{\Lambda}^{-2}(W_i - \boldsymbol{D}\boldsymbol{\mu}_{y_i})$. The rest of the sampling steps are the same.

Update $\boldsymbol{Y}$. The cause-of-death assignment can be updated by calculating the posterior probability of belonging to each cause by $\Pr(Y_i = c|\boldsymbol{Z}_i, \boldsymbol{\mu}, \tilde{\boldsymbol{R}}) \propto \phi(\boldsymbol{Z}_i; \boldsymbol{\mu}_c, \tilde{\boldsymbol{R}})$.

Update $\boldsymbol{\pi}$. The update of the CSMF follows similar to the algorithm in McCormick et al. (2016). We first sample the latent mean and variance by

$$\mu_\theta \sim \text{Normal}(\frac{1}{C}\sum_c \theta_c, \frac{\sigma_\theta^2}{C}),$$

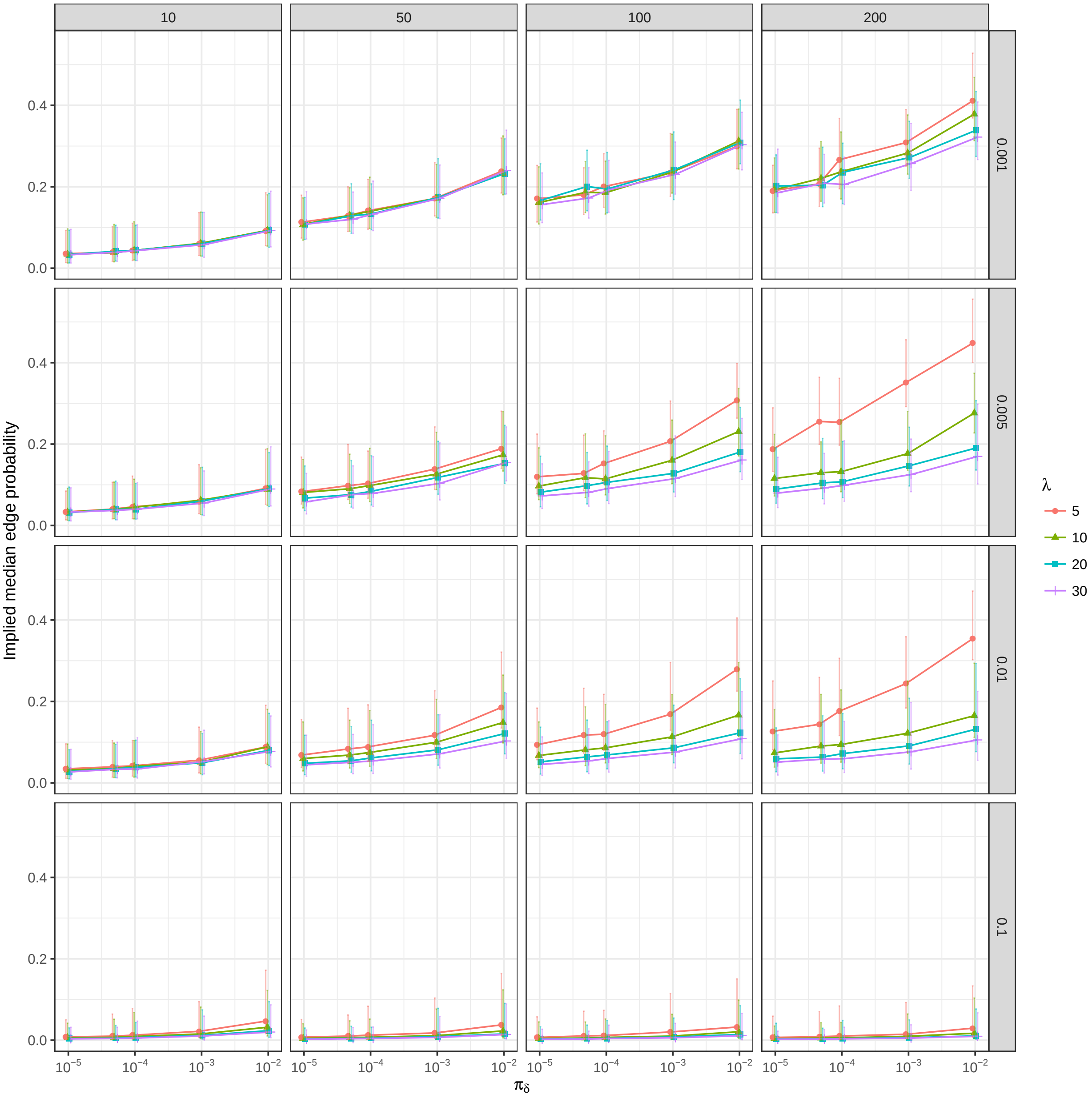


Figure 3: Implied prior edge probability for $p = 100$ graph. The dots represent the median prior probabilities and the error bars represent the 0.025 and 0.975 quantiles The rows in the panel represent the value of v_0, and the columns represent the choice of v_1/v_0. For each combination of v_0 and v_1, the edge probabilities induced by different λ and $\pi_{\boldsymbol{\delta}}$ are plotted. The densities are derived from sampling $1,000$ draws using MCMC from the prior distribution after $1,000$ iterations of burn-in.

$$\sigma_\theta^2 \sim \text{Inv-}\chi^2(C-1, \frac{1}{C}\sum_c (\theta_c - \mu_\theta)^2).$$

Then we sample $\boldsymbol{\theta}|n_c, \mu_\theta, \sigma_\theta^2 \propto prod_c \pi_c^{n_c}\text{Normal}(\boldsymbol{\theta}; \mu_\theta, \sigma_\theta^2\boldsymbol{I})$ using ESS, where n_c is the number of deaths assigned to cause c.

Update σ_c^2. When σ_c^2 is not fixed in the model, we can sample them from the conjugate posterior distribution

$$\sigma_c^2 \sim \text{InvGamma}(0.001 + \frac{p}{2}, 0.001 + \frac{\sum_{j=1}^{p}(\mu_{cj} - \mu_{0cj})^2}{2})\ .$$

5 Evaluation of the Gaussian approximation in Section 3.1

In Section 3.1 of the main paper, the posterior samples of v was taken using with a Gaussian approximation step of the conditional density. That is, we approximate the true conditional distribution

$$p(v|\boldsymbol{u}, \boldsymbol{S}, \boldsymbol{V}) \propto \text{Gamma}(v; \frac{n}{2}, \frac{s_{jj}+1}{2}) \exp\left(-\frac{1}{2v}\boldsymbol{u}'(\hat{\boldsymbol{D}} + \tilde{\boldsymbol{D}}(\boldsymbol{u}, v) + \lambda\boldsymbol{\Omega}^{-1}_{[-j,-j]})\boldsymbol{u}\right)$$

with

$$p(v|\boldsymbol{u}, \boldsymbol{S}, \boldsymbol{V}) \propto \text{Normal}(v; \frac{n}{s_{jj}+1}, \frac{2n}{(s_{jj}+1)^2}) \exp\left(-\frac{1}{2v}\boldsymbol{u}'(\hat{\boldsymbol{D}} + \tilde{\boldsymbol{D}}(\boldsymbol{u}, v) + \lambda\boldsymbol{\Omega}^{-1}_{[-j,-j]})\boldsymbol{u}\right)$$

since s_{jj} is typically much smaller than n which makes the Gamma density well approximated by the Normal distribution. To assess this approximation, we additionally implemented a modified ESS approach by rewriting the correct conditional density into

$$p(v|\boldsymbol{u}, \boldsymbol{S}, \boldsymbol{V}) \propto \text{Normal}(v; \frac{n}{s_{jj}+1}, \frac{2n}{(s_{jj}+1)^2}) \exp\left(-\frac{1}{2v}\boldsymbol{u}'(\hat{\boldsymbol{D}} + \tilde{\boldsymbol{D}}(\boldsymbol{u}, v) + \lambda\boldsymbol{\Omega}^{-1}_{[-j,-j]})\boldsymbol{u}\right) R(v; s_{jj}, n)$$

where $R(v; s_{jj}, n)$ is the ratio between the Gamma and Normal densities. Notice that this approach allows the exact likelihood to be sampled at each step, but could potentially suffer from slow mixing (Nishihara et al., 2014). The approximation used in the paper leads to very similar posterior means of the parameters compared to sampling from this exact likelihood. Figure 4 shows the comparison of the posterior means of $\boldsymbol{R}$ and $\boldsymbol{\mu}$ using the two sampling schemes. The posterior means obtained from the approximation shrink slightly more to zero but with good agreement to the ones drawn from the exact likelihood.

6 Additional simulation evidence of classification accuracies

6.1 Classification error

In this section we illustrate the performance of our method for cause-of-death assignment in VA analysis. We generate $n = 800$ unlabeled data with $p = 50$ from $C = 20$ classes, where the class membership distributions are generated from Dirichlet(1). Data within all groups share the same latent correlation matrix but have different marginal mean vectors generated in the same way as described in the main paper.

For the proposed model, we further investigate the scenario where 0, 100 and 200 labeled data exist. Intuitively, adding labeled data helps our model identify the dependence structure more quickly, especially in the presence of low sample size and high proportion of missing data. However, we do not impose the assumption that the labeled data shares the same class distribution as the testing data to maintain fair comparison. Figure 5 and 6 display the results in terms of the CSMF accuracy and classification accuracy. The proposed latent Gaussian model consistently outperforms both the naive Bayes classifier and InterVA model, and is more robust to misspecification.

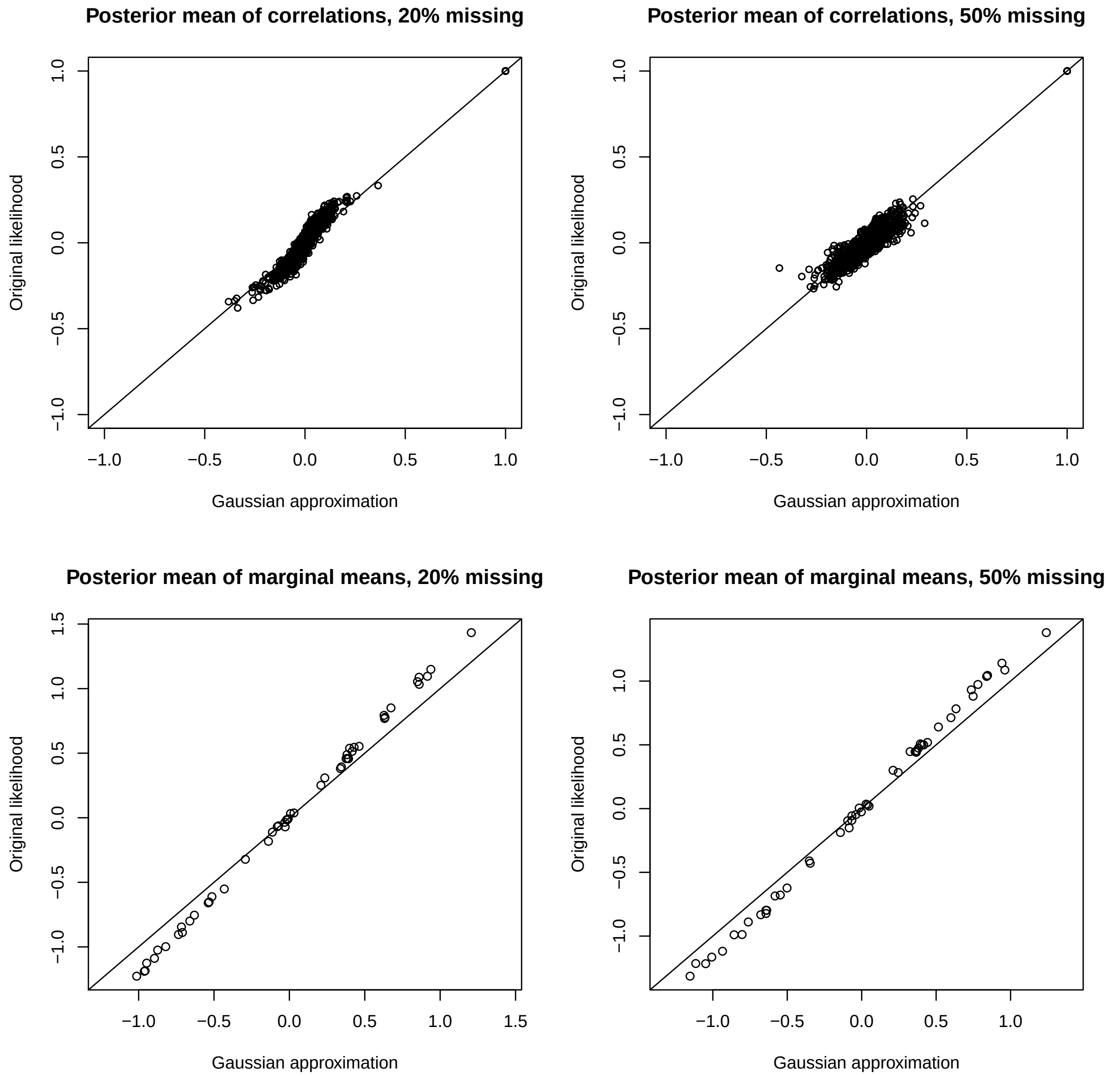


Figure 4: Compare the estimated latent correlation matrix and latent marginal means using approximate and exact likelihood. The data are simulated as in Case (ii) described in the main paper with 20% and 50% of missing data respectively. Both samplers are run $10,000$ iterations with the first half discarded and every 10th iteration saved.

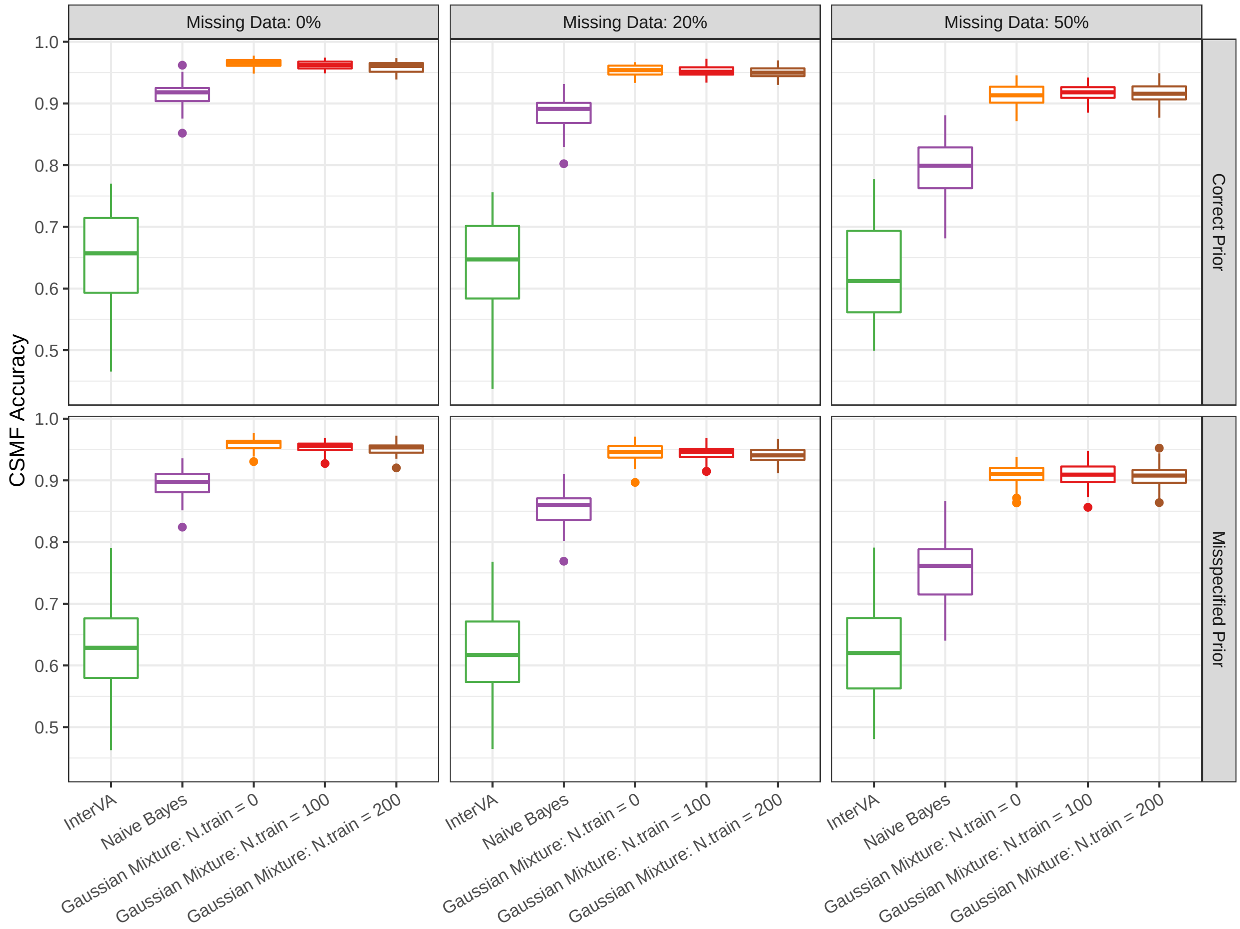


Figure 5: Box plot of CSMF accuracy for simulated mixed data. The accuracy is evaluated in a dataset with a total $n = 800$ observations and $p = 50$ variables including 5 continuous variables from $C = 20$ classes, under both correct and misspecified priors and different proportion of missing data.

7 Convergence analysis

7.1 Examples with simulated data

In the simulation analysis with $p = 50$ and a single class, the posterior draws converge fairly quickly. Figure 7 shows the trace plots of the graph size and a random selection of mu_j from 5 chains with different starting values in a single simulation with misspecified priors and 20% missing data. The Gelman-Rubin statistics for the graph size and $\boldsymbol{\mu}$ are all less than 1.01.

7.2 Examples with VA data

In this section we present the Gelman-Rubin statistics for fitting the proposed model to the Karonga data. We focus on the convergence of CSMF. We ran the Karonga data with four chains from different starting values. We drew Table 1 shows the Gelman-Rubin statistics for the CSMF vector ordered by the prevalence. The statistics are mostly close to 1 except for causes with small fractions. Similar difficulties in convergence of small CSMFs have been previously reported in McCormick et al. (2016) as well. The traceplots in Figure 8 and 9 show that the CSMFs converge to the same levels from multiple chains.

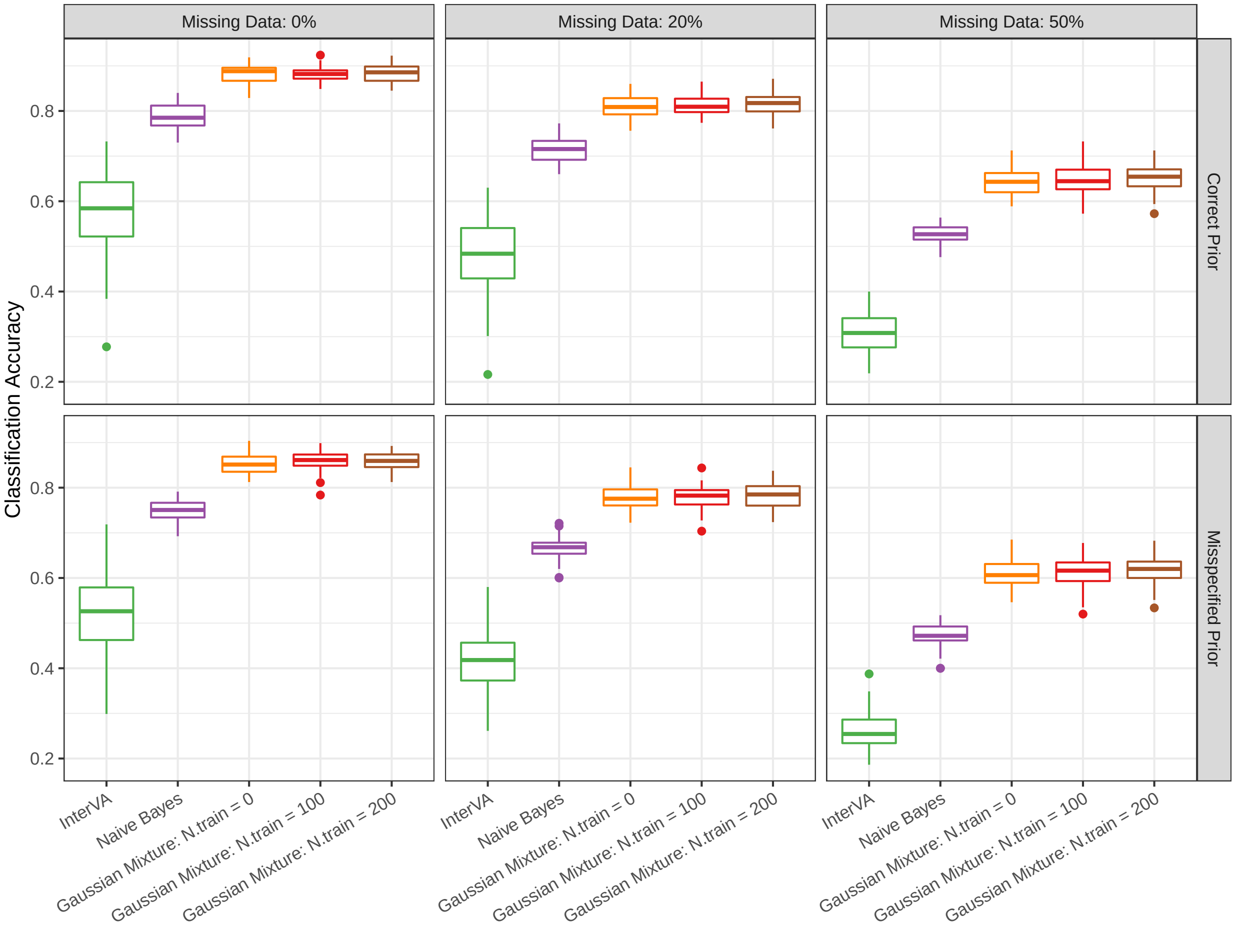


Figure 6: Box plot of classification accuracy for simulated mixed data. The accuracy is evaluated in a dataset with a total $n = 800$ observations and $p = 50$ variables including 5 continuous variables from $C = 20$ classes, under both correct and misspecified priors and different proportion of missing data.

8 More details about the Karonga data analysis

8.1 Distribution of causes of death in Karonga data

A figure representation of the causes-of-death distribution in the Karonga dataset used in the experiments are presented in Figure 10.

8.2 Estimated dependence structures

In this subsection, we include some additional results of the analysis in Section 6.2 of the main paper using pre-2008 data as training set and all the rest as testing data. The estimated correlation matrix, inverse correlation matrix, and the posterior inclusion probabilities of edges are shown in Figure 11.

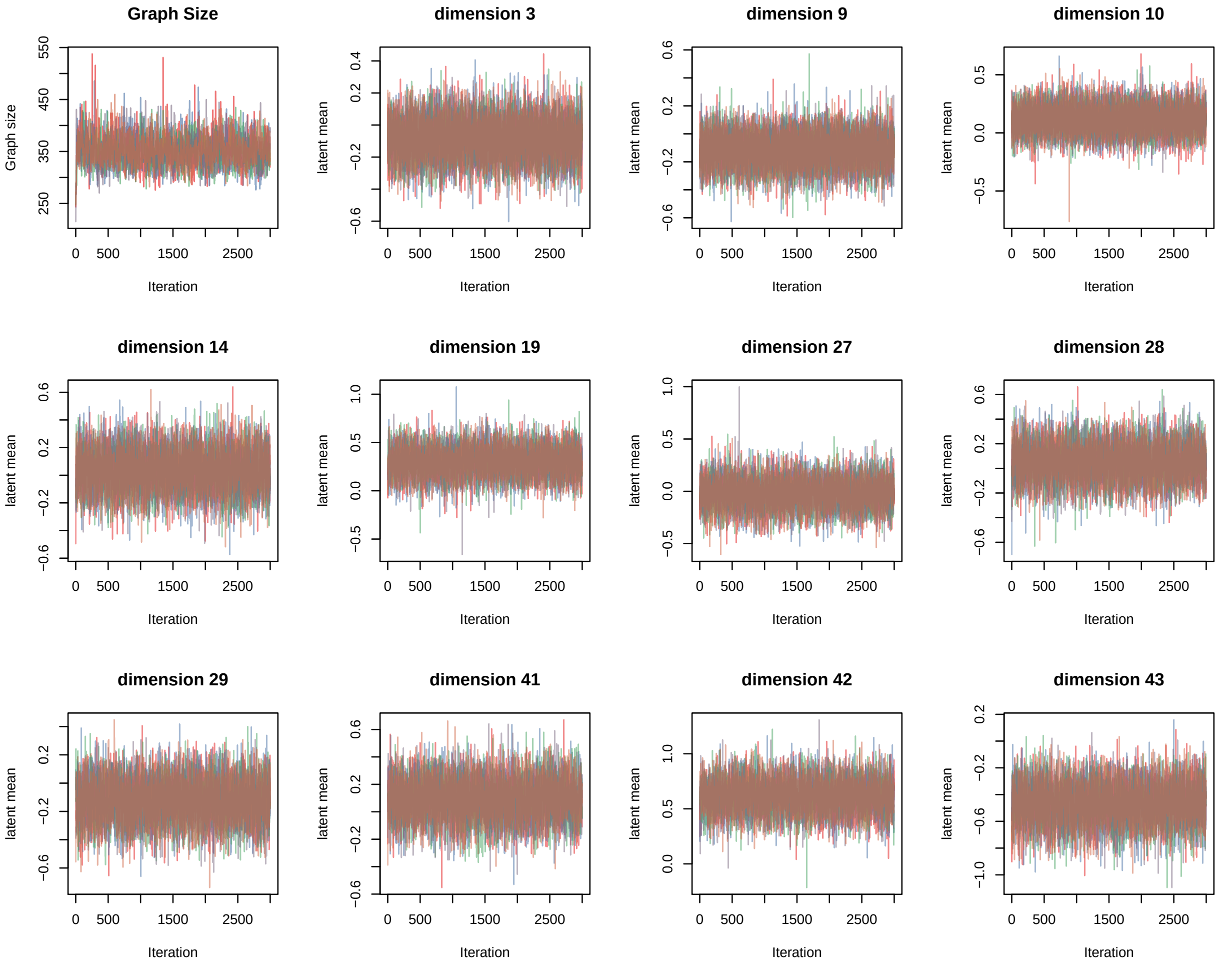


Figure 7: Trace plots of the graph size and a random subset of the marginal means μ_j. The colors indicate five chains with different starting values.

9 Numerical illustration of structural bias of independence assumption in VA analysis

In this section, we provide a numerical illustration to show the influence of ignoring correlation in cause-of-death assignment. We note that similar ideas of incorporating the dependencies between predictors for prediction have been studied recently in regression analysis (e.g. Guan et al., 2016; Peterson et al., 2015). For Naive Bayes classification, many previous studies have shown that it is, in many scenarios, robust to ignored dependencies (e.g., Rish, 2001), yet we are not aware of any formal discussion of the independence assumption in VA analysis. Here we illustrate some potential issues with the following example.

Assume the simple scenario where only three infectious diseases $C = (c_1, c_2, c_3)$ are of interest. For example, HIV/AIDS, TB, and a third category of "undetermined infectious disease", which in general includes deaths possibly due to either HIV/AIDS or TB but cannot be determined from data. Assuming there are two symptoms $S = (s_1, s_2)$, and denoting $P_{s_1 s_2}(C) = \Pr(C|S = (s_1, s_2))$, $p_i = \Pr(s_1 = 1|C_i)$ and $q_i = \Pr(s_2 = 1|C_i)$, we can write the conditional distribution for the four combinations of S as follows under

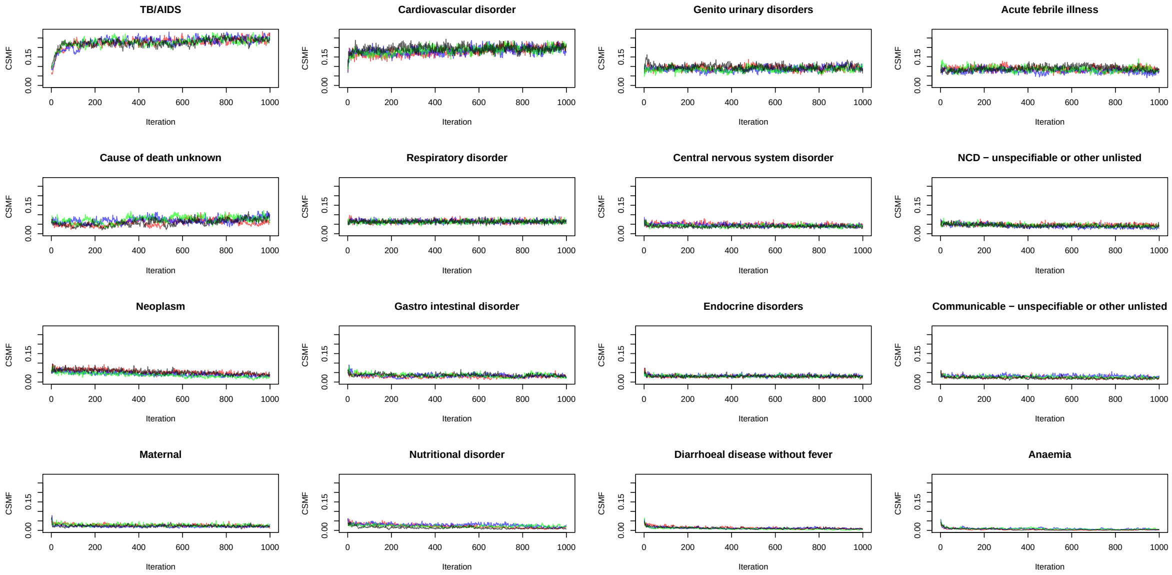


Figure 8: Trace plots for each CSMF posterior. Samples from four chains including the burn-in period, arranged in descending order by the mean.

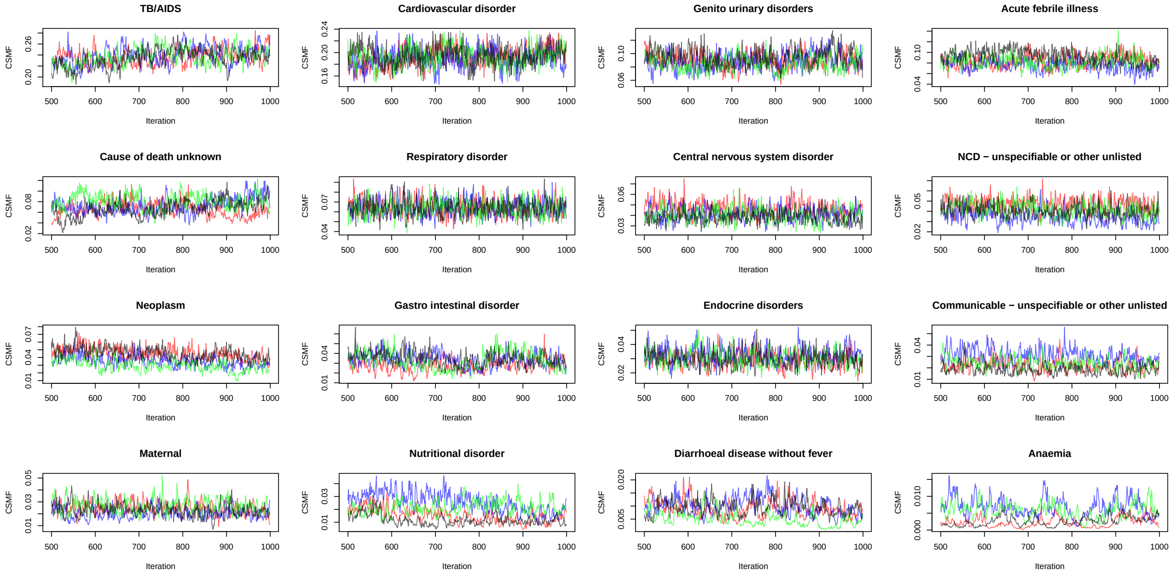
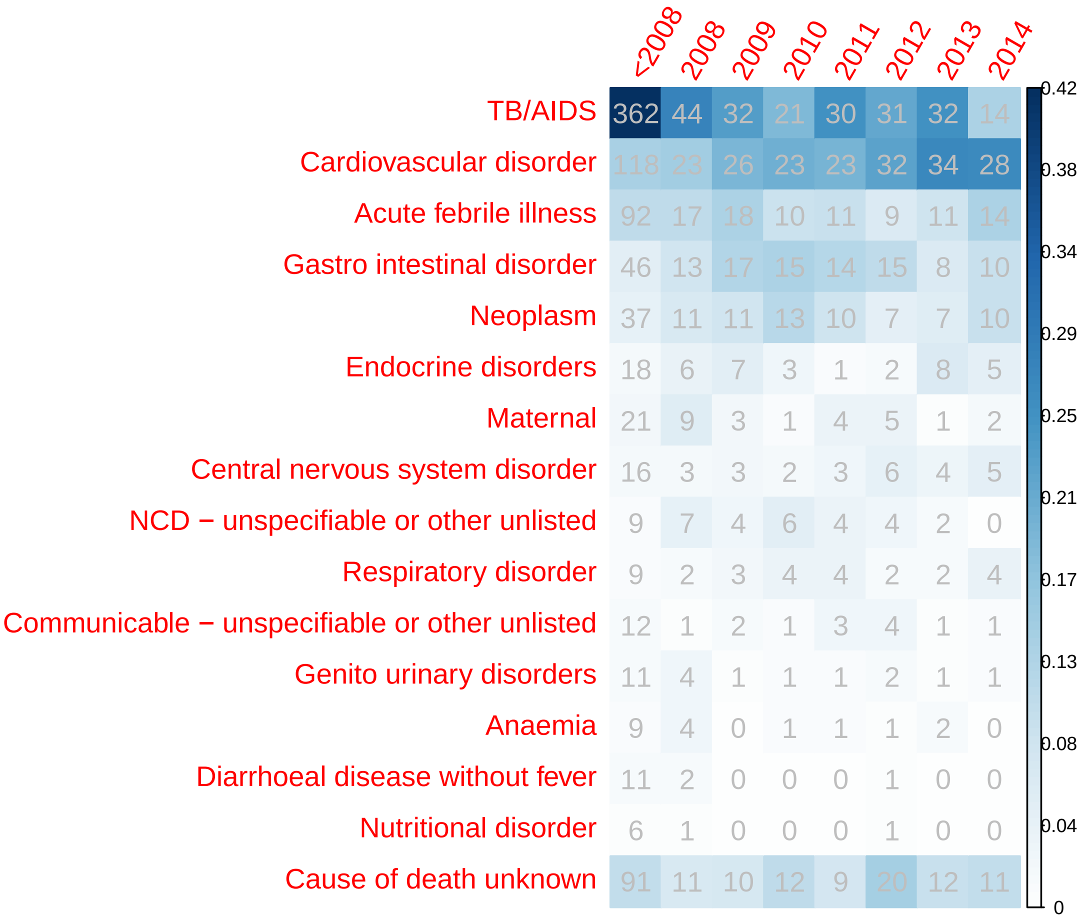


Figure 9: Trace plots for each CSMF posterior. Samples from four chains after the burn-in period, arranged in descending order by the mean.



	<2008	2008	2009	2010	2011	2012	2013	2014
TB/AIDS	362	44	32	21	30	31	32	14
Cardiovascular disorder	118	23	26	23	23	32	34	28
Acute febrile illness	92	17	18	10	11	9	11	14
Gastro intestinal disorder	46	13	17	15	14	15	8	10
Neoplasm	37	11	11	13	10	7	7	10
Endocrine disorders	18	6	7	3	1	2	8	5
Maternal	21	9	3	1	4	5	1	2
Central nervous system disorder	16	3	3	2	3	6	4	5
NCD − unspecifiable or other unlisted	9	7	4	6	4	4	2	0
Respiratory disorder	9	2	3	4	4	2	2	4
Communicable − unspecifiable or other unlisted	12	1	2	1	3	4	1	1
Genito urinary disorders	11	4	1	1	1	2	1	1
Anaemia	9	4	0	1	1	1	2	0
Diarrhoeal disease without fever	11	2	0	0	0	1	0	0
Nutritional disorder	6	1	0	0	0	1	0	0
Cause of death unknown	91	11	10	12	9	20	12	11

Figure 10: Distributions of causes-of-death in Karonga dataset by year. The integers in each cell show the number of deaths in the corresponding period, and the shading represents the proportion of causes in each year. The data before 2008 are used as prior information in the experiment and thus are combined in this figure.

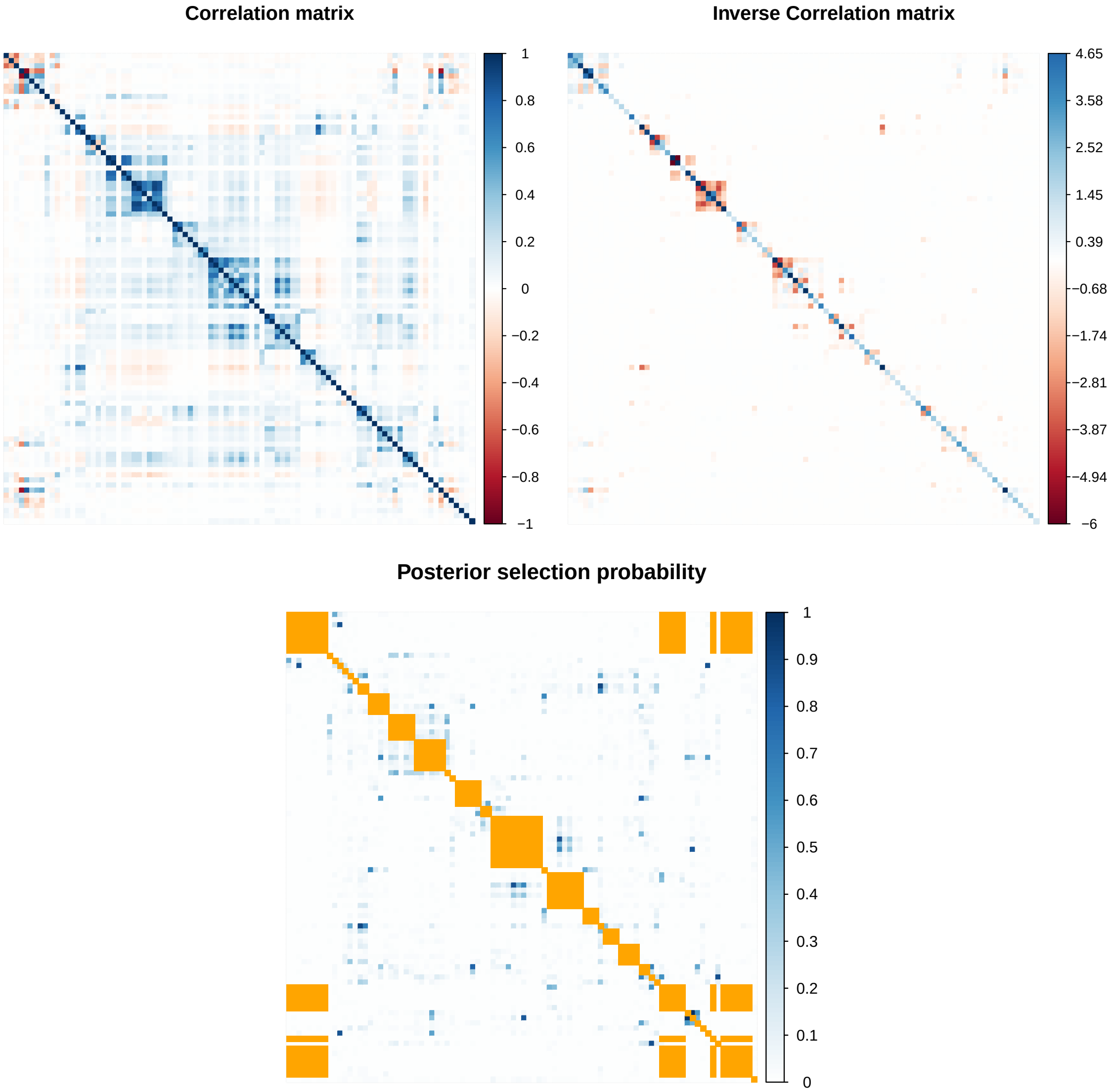


Figure 11: Posterior mean correlation (upper left), inverse correlation (upper right), and the inclusion probability (lower) matrix for Karonga data. The cells with orange color are the known edges from the questionnaire structure that is not estimated.

	Mean	$RHat$
Graph size	474.59	1.02
TB/AIDS	0.25	1.07
Cardiovascular disorder	0.20	1.05
Acute febrile illness	0.09	1.05
Genito urinary disorders	0.09	1.08
Cause of death unknown	0.08	1.08
Respiratory disorder	0.06	1.04
Central nervous system disorder	0.04	1.10
NCD - unspecifiable or other unlisted	0.04	1.25
Gastro intestinal disorder	0.03	1.05
Endocrine disorders	0.03	1.05
Neoplasm	0.03	1.41
Communicable - unspecifiable or other unlisted	0.02	1.34
Maternal	0.02	1.38
Nutritional disorder	0.01	1.66
Diarrhoeal disease without fever	0.00	1.23
Anaemia	0.00	1.41

Table 1: Gelman-Rubin statistics for graph size and CSMF in the Karonga example.

the independence assumption

$$Pr(S|C_i) = \begin{array}{cc} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{pmatrix} (1-p_i)(1-q_i) & (1-p_i)q_i \\ p_i(1-q_i) & p_i q_i \end{pmatrix} \end{array} \qquad i = 1,2,3$$

Applying Bayes rule with uniform prior on the prior distribution of the three causes of death, we can see the entries in the table above are proportional to the posterior probability of assigning each cause of death given a specific observation of symptoms, since

$$P_{s_1,s_2}(C_i) = \frac{\frac{1}{3}P(S|C_i)}{\sum_{j=1}^{3}\frac{1}{3}P(S|C_j)} = \frac{P(S|C_i)}{\sum_{j=1}^{3}P(S|C_j)} \propto P(S|C_i) \ .$$

Now consider the case where the two symptoms s_1 and s_2 are respectively key symptoms for c_1 and c_2, so that $p_1 > p_2$ and $q_1 < q_2$. Since deaths due to c_3 are essentially a mixture of the other two causes and we assume equal prevalence of c_1 and c_2, we can roughly let $P(S|C_3) = P(S|C_1)/2 + P(S|C_2)/2$. Still using the independence assumption for c_1 and c_2, we calculate the correct joint distribution of symptoms given c_3 to be

$$\begin{aligned} Pr(S|C_3) &= \begin{array}{cc} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{pmatrix} \theta_{00} & \theta_{10} \\ \theta_{01} & \theta_{11} \end{pmatrix} \end{array} \\ &= \begin{array}{cc} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{pmatrix} ((1-p_1)(1-q_1)+(1-p_2)(1-q_2))/2 & ((1-p_1)q_1+(1-p_2)q_2)/2 \\ (p_1(1-q_1)+p_2(1-q_2))/2 & (p_1q_1+p_2q_2)/2 \end{pmatrix} \end{array} \end{aligned}$$

which violates the independence assumption since the product of marginal probabilities $\Pr(s_1 = 1|C_3)\Pr(s_2 = 1|C_3) = (\theta_{10}+\theta_{11})(\theta_{01}+\theta_{11}) = (q_1+q_2)(p_1+p2)/4 > (p_1q_1+p_2q_2)/2 = \theta_{11}$ when $(p_1-p_2)(q_1-q_2) < 0$. This implies that by naively applying Bayes rule and assuming independence of symptoms, we will over-estimate $P_{11}(C_3)$ under this setup.

Additionally, we consider the scenario where $p_1 = q_2$ and $q_1 = p_2$, which is highly likely when the conditional probabilities are provided as rankings instead of numerical values, as in the implementation of

InterVA. It is obvious to show that $\Pr(s_1 = 1|C_3) \Pr(s_2 = 1|C_3) = (q_1 + q_2)(p_1 + p_2)/4 = (q_1 + p_1)^2/4 > q_1 p_1$, which means if independence of symptoms conditional on causes is assumed, a researcher will conclude $P_{11}(C_3) > P_{11}(C_1)$, and similarly $P_{11}(C_3) > P_{11}(C_2)$. In contrast, if the analysis is carried out with the correct conditional probability table, it should lead to $P_{11}(C_1) = P_{11}(C_2) = P_{11}(C_3)$ since the lower right entries in all three tables are equal. This heuristic example shows that even when some of the conditional independence assumptions are satisfied and all marginal $P_{s|c}$ are accurately estimated, due to the particular features of VA analysis that includes causes that are "undetermined", the independence assumption can lead to undesired outcomes that overestimate the "undetermined" categories. These biases *result entirely from model assumptions* and cannot be solved with more data, and the problem becomes even worse as the number of symptoms and causes grows.

References

Guan, L., Fan, Z., and Tibshirani, R. (2016). "Regularization for supervised learning via the " hubNe" procedure." *arXiv preprint arXiv:1608.05465*. 10

McCormick, T. H., Li, Z. R., Calvert, C., Crampin, A. C., Kahn, K., and Clark, S. J. (2016). "Probabilistic cause-of-death assignment using verbal autopsies." *Journal of the American Statistical Association*, 111(515): 1036–1049. 4, 8

Nishihara, R., Murray, I., and Adams, R. P. (2014). "Parallel MCMC with generalized elliptical slice sampling." *The Journal of Machine Learning Research*, 15(1): 2087–2112. 6

Peterson, C. B., Stingo, F. C., and Vannucci, M. (2015). "Joint Bayesian variable and graph selection for regression models with network-structured predictors." *Statistics in Medicine*, (October). 10

Rish, I. (2001). "An empirical study of the naive Bayes classifier." In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, 41–46. IBM New York. 10

Wang, H. (2015). "Scaling it up: Stochastic search structure learning in graphical models." *Bayesian Analysis*, 10(2): 351–377. 2