CoDCorrect: Validation of a latent dirichlet model via simulation study

Abraham D. Flaxman*[1] and Laura Dwyer-Lindgren*[1]

[1]Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA

Email: Abraham D. Flaxman*- abie@uw.edu; Laura Dwyer-Lindgren*- ladwyer@uw.edu;

*Corresponding author

Abstract

Background: Population-level information on cause-specific mortality rates by country, age, sex, and time is important for decision-making. Models are crucial to estimating and predicting this information when direct measurements are not available. State-of-the-art models produce estimates disease by disease, however, and do not enforce the constraint that the sum of deaths over every cause must be equal to the total deaths due to all causes.

Methods: We developed a statistical model to correct for inconsistencies between cause-specific mortality rate estimates and all-cause mortality estimates. We compared our model to a naive approach in a validation environment using a series of simulation studies designed to test TK. We measured the quality of the corrected estimates by TK.

Results: Over all simulations, our approach performed TK, compared to TK for the naive approach. For setting TK, our approach showed significant benefits in terms of TK and TK. For setting TK, neither approach performed very well, demonstrating that without sufficiently accurate input data, all model-based approaches have limitations.

Conclusions: Our model is a fast and effective way to improve the estimates produced by many separate models, and we recommend using it to enforce consistency between cause-specific mortality rates and all-cause mortality rates.

Background

TK one paragraph on population health metrics, GBD, CoD paper which uses different ensemble models for each cause and how hard it would be to combine them.

Models for each cause are estimated separately, so there is no guarantee of consistency and the estimated cause fractions for any given country-sex-year do not necessarily sum to one.

TK one paragraph on what we have done, a model-based approach to enforcing consistency between cause-specific model estimates and all-cause estimates. We validated our approach through a simulation study, with an environment constructed to capture important characteristics of the real challenge. We propose two models that take different approaches to predicting consistent cause fractions from inconsistent preliminary estimates. These two models are then validated in the simulation environment and compared on a variety of metrics.

TK one paragraph on the importance of what we have done.

Methods

The estimation challenge we faced was the following: For a given sex, age group, and geographic region, we obtained estimates of the cause-specific mortality fraction F_j for a mutually exclusive, collectively exhaustive list causes $j = 1, \ldots, J$, using a statistical model described elsewhere **??**. Each estimate is represented by N draws from the posterior distribution over values for T years:

$$F_j = \left(f^n_{j,y_1}, \ldots, f^n_{j,y_T}\right)_{n=1}^{N}.$$

Below we have also used the notation $F_{j,y}$ to denote the marginal empirical distribution of F_j for year y:

$$F_{j,y} = \left(f^1_{j,y}, \ldots, f^N_{j,y}\right).$$

Our goal was to produce a posterior distribution for the true cause fractions $\pi_{j,y}$, by incorporating the consistency condition $\sum_j \pi_{j,y} = 1$ for all y. This would allow us to produce point estimates and uncertainty intervals for individual cause fractions, as well as estimate the change in $\pi_{j,y}$ as a function of y for each j.

Quality Metrics

We wanted the estimated cause fractions from any model to be close to the true cause fractions in terms of absolute error, although being accurate in terms of relative error is also important for some applications. Following the approach taken in Verbal Autopsy quality metrics **??**, we summarize the absolute error using cause specific mortality fraction accuracy (CSMF accuracy), defined by the following formula:

$$\text{CSMF Accuracy} = 1 - \sum \frac{\left|\pi_{j,y} - \pi_{j,y}^{\text{true}}\right|}{2(1 - \min_j(\pi_{j,y}^{\text{true}}))}$$

We also wanted the estimates to have accurate uncertainty intervals, which we quantified in terms of coverage probability (the probability that the true estimate falls between the upper and lower bounds of the 95% highest probability density (HPD) interval). We desired that the coverage probability for each cause be close to 0.95.

TK description of bias, another important metric of estimation quality:

$$\text{Median Bias} = \text{Median}\, j \left\{\pi_{j,y} = \pi_{j,y}^{\text{true}}\right\}$$

TK description of trend accuracy, a newly developed test statistic, designed to capture how accurately our model has captured the time trend of the corrected cause fraction estimates:

$$\text{Trend Accuracy} = TK$$

Statistical Models

Before developing a sophisticated statistical model for estimating $\pi_{j,y}$, we considered a naive approach: for each $n = 1, \ldots, N$, for each y, we simply scaled the vector $(f_{1,y}^n, \ldots, f_{J,y}^n)$ to make it sum to unity:

$$\pi_{j,y}^n = \frac{f_{j,y}^n}{\sum_{j'=1}^J f_{j',y}^n}.$$

This provided a non-parametric posterior distribution for $\pi_{j,y}$, which we summarized as a point estimate for each j and y by taking the mean over values of n, and estimates uncertainty intervals by taking the 95% HPD interval.

We developed the more traditional statistical model that we now describe to take into account the differences in uncertainty that the different $f_{j,y}$ in the input data often exhibit. The formal specification of the model is the following:

$$\begin{aligned}
\operatorname{dens}(F_{j,y}|\pi) &\propto \exp\left\{-\frac{(\pi_{j,y} - \mathrm{E}[F_{j,y}])^2}{\mathrm{Var}[F_{j,y}]}\right\}\\
\pi_{j,y} &= \frac{\alpha_{j,y}^t}{\sum_{j'=1}^J \alpha_{j',y}^t}\\
\alpha_{j,y} &\sim \text{Normal}\left(0, 1^2\right)
\end{aligned}$$

Here the notation $\mathrm{E}[F_{j,y}]$ and $\mathrm{Var}[F_{j,y}]$ are the usual mean and variance:

$$\begin{aligned}
\mathrm{E}[F_{j,y}] &= \left(\sum_{n=1}^N f_{j,y}^n\right)/N,\\
\mathrm{Var}[F_{j,y}] &= \left(\sum_{n=1}^N \left(f_{j,y}^n - \mathrm{E}[F_{j,y}]\right)^2\right)/N.
\end{aligned}$$

This approach can be modified to explicitly account for time-correlation in the F_j distribution, by making the likelihood look more like a multivariate normal:

$$\begin{aligned}
\operatorname{dens}\left(F_j|\pi\right) &\propto \exp\left\{-\left(\pi_j - \mathrm{E}[F_j]\right)^T \Sigma_j^{-1}\left(\pi_j - \mathrm{E}[F_j]\right)\right\}\\
\pi_{j,y} &= \frac{\alpha_{j,y}^t}{\sum_{j'=1}^J \alpha_{j',y}^t}\\
\alpha_{j,y} &\sim \text{Normal}\left(0, 1^2\right)\\
\Sigma_j &= \text{Matern}(\sigma, \rho_j, 2) + \text{Diag}(\text{Var}[F_j]))\sigma \qquad \sim \text{Uniform}[0,1]\\
\rho_j &\sim \text{Normal}_{\epsilon+}(20, 10^2)
\end{aligned}$$

Here $\mathrm{E}[F_j]$ and $\mathrm{Var}[F_j]$ are the vector-valued analogues of the expectation and variance above, and Matern $= TK$. TK more description of the Matern covariance.

Simulation Environment

To validate the models above, and to compare the sophisticated approaches to the naive approach, and to generally assess how much improvement this sort of consistenty constraint can be expected to provide, we conducted a simulation study. We chose true starting cause fractions π^{true} for a list of 10 causes from a heavy-tailed distribution TK, and then randomly perturbed them to vary smoothly over a 10 year time period, according to a Matern covariance function TK. Then we used simulation to generate 1000 samples of noisy "estimates" of the true cause fractions to produce simulated input data for which ground truth was known. We varied the dispersion of the noisy estimates, as well as the bias to understand how the different models perform in different settings. We also varied the correlation between the bias and the variance, to explore how inappropriately narrow uncertainty in the CoD model would affect the corrected estimates. This can be understood as a three-armed factorial study design, with dispersion taking values low, medium, and high, and bias taking values low, medium, and high, and correlation taking values of positive, zero, and negative.

To be precise, the true mean cause fractions were generated by a normalized exponential transform of 10 Gaussian Processes with covariance Matern(TK), evaluated at times $1, 2, \dots, 10$:

$$\begin{aligned} \alpha_j^{\text{true}}(y) &\sim \text{GP}(0, \text{Matern}(TK)), \\ \pi_{j,y}^{\text{true}} &= \frac{\alpha_j(y)}{\sum_{j'=1}^{J} \alpha_{j'}(y)}. \end{aligned}$$

Then estimate of cause fractions were generated for draw n from a two stage perturbation as follows:

$$\begin{aligned} \text{logit}\, f_{j,y}^{n} &= \text{logit}(\pi_{j,y}^{\text{true}} + \beta_{j,y} + \epsilon_{j,y,n}), \\ \beta_{j,y} &\sim \text{Normal}(0, \sigma_{\text{bias}}^2), \\ \epsilon_{j,y,n} | \beta_{j,y} &\sim \text{Normal}(TK, \Sigma^2), \end{aligned}$$

$$\Sigma^2 = TK.matrix.with \sigma_{\text{bias}}^2, \sigma_{\text{dispersion}}^2, \rho_{\text{correlation}}.$$

This results in N draws from a distribution with the same bias and dispersion.

We repeated this procedure 1000 times for each of the 27 possibilities in the design, and calculated CSMF accuracy, coverage probability, and TK something about change over time for all of them.

Results/Discussion

Results sub-heading

Another results sub-heading

Yet another results sub-heading

Conclusions

The good model is better than the bad model! Hurray!!!

List of abbreviations used

Competing interests

Authors contributions

Authors information

Acknowledgements and Funding

Figures

Figure 1 - Sample figure title

A short description of the figure content should go here.

Figure 2 - Sample figure title

Figure legend text.

Tables

Table 1 - Sample table title

Here is an example of a *small* table in LaTeX using `\tabular{...}`. This is where the description of the table should go.

My Table		
A1	B2	C3
A2	...	..
A3	..	.

Table 2 - Sample table title

Large tables are attached as separate files but should still be described here.

Additional Files

Additional file 1 — Sample additional file title

Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title

Additional file descriptions text.