Optimizing Influenza Vaccine Composition: From Predictions to Prescriptions

Hari Bandi HBANDI@MIT.EDU
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139

Dimitris Bertsimas DBERTSIM@MIT.EDU
Sloan School of Management and Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

We propose a holistic framework based on state-of-the-art methods in Machine Learning and Optimization to prescribe influenza vaccine composition that are specific to a region, or a country based on historical data concerning the rates of circulation of predominant viruses. First, we develop a tensor completion formulation to predict rates of circulation of viruses for the next season based on historical data. Then, taking into account the uncertainty in the predicted rates of circulation of predominant viruses, we propose a novel robust prescriptive framework for selecting suitable strains for each subtypes of the flu virus: Influenza A (H1N1 and H3N2) and B viruses for production. Finally, we train optimal regression trees to predict efficacy of the prescribed vaccine in terms of both morbidity and mortality rates using a set of weighted distances between the vaccine-strain and the actual circulating viruses during a flu season for each subtypes of the flu virus. Through numerical experiments, we show that our proposed vaccine compositions could potentially lower morbidity by 11-14% and mortality by 8-11% over vaccine compositions proposed by World Health Organization (WHO) for Northern hemisphere, and lower morbidity by 8-10% and mortality by 6-9% over vaccine compositions proposed by U.S Food and Drug Administration (FDA) for USA, and finally, lower morbidity by 10-12% and mortality by 9-11% over vaccine compositions proposed by European Medicines Agency (EMA) for Europe.

1. Introduction

Influenza (flu) is a highly contagious respiratory viral disease and the seasonal flu epidemics affect about 5-15% of the world's population, and cause an estimated 3-5 million cases of severe illnesses and up to half a million annual deaths worldwide. The flu viruses can be segregated into four types, namely influenza A, B, C and D. Influenza A and B viruses circulate and therefore are mainly responsible for seasonal flu epidemics, whereas influenza C viruses are not detected frequently and usually cause mild infections; influenza D viruses affect cattle and thus do not present a serious public health risk. Influenza A viruses are classified on the basis of their two surface proteins: hemagglutinin and neuraminidase, into 18 different subtypes of hemaglutinin and 11 different subtypes of neuraminidase viruses.

Together, all subtypes of influenza A and influenza B viruses are further classified into various strains based on their *antigenic* properties (response to antibodies).

The flu shot (vaccine), which contains two strains of the influenza A virus (H1N1 and H3N2) and one or two strains of the B virus is the first line of defense against seasonal epidemics. The influenza B viruses have two distinct lineages, therefore in addition to a vaccine containing three virus-strains, manufacturers also produce a vaccine containing four virus-strains vaccine which includes two influenza B strains to cover both lineages. Most individuals have some level of prior immunity. However, new strains with mutations in their protein regions that are not recognized by human antibodies frequently arise and it is well established that these strains have an advantage over existing dominant strains to effectively escape from host immunity. This continuous process of evolution results in a rapid turnover of the viral population and poses a great challenge to producing an effective vaccine.

The flu vaccine is annually reformulated due to rapid emergence of new strains, and prepared at least six to eight months in advance of the upcoming flu season in order to have enough time for production and distribution. Every year, the vaccine compositions for the Northern and Southern Hemispheres are reviewed and updated as necessary by the World Health Organization (WHO) through a global surveillance and response system. The WHO monitors and collects data on antigenic characterization, genetic variations, prevalence rates, and geographic distributions of virus variants across the world. Although, antigenic characterization of circulating viruses by standard ferret antibodies is the main determinant in vaccine strain selection, many approaches have been proposed to partially explain mutations in the virus strains using genomic data in the literature. Predicting emerging virus strains is a complex problem due to the uncertain nature of the continuous evolution of the virus strains. Moreover, predicting the fate of strains currently circulating in the population is difficult for two reasons. First, multiple strains carrying different combinations of mutations co-circulate and therefore compete with one another for potential hosts. Second, antigenic characterization via ferret antibodies is different from that of human post-vaccination antibodies because humans and ferrets have different immune systems as well as very different prior exposure to influenza viruses (Agor and Özaltın, 2018).

In order to tackle these challenges, we employ a variety of state-of-the-art methods in machine learning and optimization to prescribe influenza vaccine composition that are specific to a region, or a country based on historical data concerning the rates of circulation of predominant viruses. Below, we briefly our outline approach.

1. **Tensor completion:** We adapt an algorithm proposed in Bertsimas and Pawlowski (2019) to predict rates of circulation of viruses in a season. The historical rates of circulation of viruses are available in the form of a three-dimensional matrix $\mathbf{M} \in \mathbb{R}^{n \times m \times T}$, where n is the number of countries participating in WHO's Global Influenza Surveillance and Response System (GISRS), m is the number of viruses in the data set and T is the number of flu seasons. Such a three-dimensional matrix is known as a three-dimensional *tensor*.

   About 88% of the tensor entries are missing. The major reason that the tensor is very sparse is that GISRS has been adding viruses to the data set over time and therefore, for newly identified viruses, we observe missing entries for all earlier flu

seasons. Moreover, as viruses keep undergoing mutations over time, not all viruses in the data set may be circulating in a particular flu season. Therefore, we observe a lot of missing entries either due to the viruses undergoing mutations or some inherent bias in the testing sample. For example, a sample tested in a particular laboratory may miss out on some emerging strains which might be crucial to identify and predict future circulating viruses.

In order to predict rates of circulation of viruses in the future, we propose to estimate a low-rank tensor to approximate the observed data such that the low-rank component of the tensor varies slowly across time. The reason that we impose such a constraint on the low-rank component of the tensor is that we assume that the weights which influence the system do not change drastically across consecutive flu seasons.

2. **Optimal regression trees:** In order to quantify the efficacy of a vaccine, we train regression trees using the Optimal Regression Trees (ORTs) algorithm proposed in Bertsimas and Dunn (2017) to predict both the morbidity and mortality rates as a percentage of the population using information about how effectively a vaccine-strain hinders a virus' ability to attack healthy red-blood cells. For this purpose, we use a distance metric also called antigenic distance between a vaccine-strain and a virus for each subtypes of the influenza virus, and define a weighted distance (see Cai et al. (2012)) between a given vaccine-strain and all predominant circulating viruses as the antigenic distance between each of these pairs weighted by the rates of circulating of the corresponding virus during a flu season.

3. **Robust optimization:** Using the same weighted distance as a metric of performance, we propose a novel robust prescriptive problem that minimizes the worst-case weighted distance given some uncertainty about the rates of circulation of viruses in the upcoming flu season in order to inform vaccine composition. The uncertainty in the rates of circulation of viruses for the upcoming flu season is quantified by restricting the low-rank component of the tensor decomposition to not deviate from its counterpart from the previous time period. We reformulate the corresponding robust prescriptive problem as a second order cone optimization probelm, and show that it is both practically and theoretically tractable.

1.1. Related work

In this section, we review some related work in literature that propose methods to model evolution of the viruses in order to inform strain selection for the seasonal influenza vaccine.

Wilson and Cox (1990) studied evolution of various virus strains and suggested that a drift variant of epidemiologic importance usually contains at least four amino acid substitutions located at more than two of the *epitope* regions (part of an antigen molecule to which an antibody attaches itself) on the HA1 polypeptide. Lee and Chen (2004) showed that the number of amino acid changes in the 131 amino acid positions around the epitope sites had the highest correlation with the antigenic distance and the best performance for predicting antigenic difference between any two virus strains. Liao et al. (2008) proposed construction of similarity classes and substitution matrices, to explain the antigenic differences of viruses' using genetic information, and employed statistical machine learning methods like iterative

filtering, multinomial logistic regression and support vector machines to quantify the effect of amino acid substitutions and identify major binding sites on the HA1 protein.

Steinbrück and McHardy (2012) used nonnegative least-squares optimization to map pairwise antigenic distances onto the branches of a *phylogenetic tree* (Neher and Bedford, 2015). Steinbrück et al. (2014) combine phylogenetic trees and *Allele Dynamics* (AD) plots to identify the HA mutations that are most likely to become predominant in the future seasons. Neher et al. (2014) proposed a method to predict the fitness of a virus from its genetic information, and analyzed the shape and branching pattern of the tree to identify the fitness of different strains relative to each other.

He and Deem (2010) construct protein distance maps for the HA1 surface proteins of the influenza 2009 A (H1N1) pandemic virus. In particular, they apply multi-dimensional scaling to project a 329-residue long amino acid sequence of the HA1 protein onto two dimensions and then use kernel density estimation to detect clusters on the protein distance map. Luksza and Lassig (2014) develop a fitness model to predict the evolution of influenza by identifying changes in the frequencies of strain groups referred to as clades. They consider two major groups of mutations at the *epitope* and *non-epitope* regions of the virus' surface protein. Mutations at *epitope* regions are likely to be beneficial to the virus, whereas, mutations outside *epitope* regions are often deleterious to the fitness of a virus. Luksza and Lassig (2014) used their model to predict frequencies of clades one year in the future with considerable accuracy. Although, some of these methods help identify a cluster or a set of viruses predicted to circulate in the future, they do not optimize the vaccine composition in anticipation of the uncertainty in their predictions.

Wu et al. (2005) formulated the strain selection problem as a stochastic dynamic program using the antigenic shape-space model proposed in Perelson and Oster (1979) where the antigenic evolution is assumed to be a random walk. Kornish and Keeney (2008) formulate the strain selection problem as a finite-horizon optimal stopping problem, where at each time step, a decision is made either to select one of two candidate strains, or to defer the selection to the next time period. Cho (2010) build upon Kornish and Keeney (2008) by considering the flu shot composition and production under yield uncertainty as a two-stage stochastic game and show the existence of an optimal threshold policy. Özaltın et al. (2018) propose a bilevel model that integrates the annual flu shot design problem of a decision maker and the profit-maximization problem of the vaccine manufacturers through a bilevel model in a stochastic environment. They model the decision maker's problem of strain selection as a two-stage stochastic mixed-integer optimization problem and propose a heuristic using Dantzig-Wolfe decomposition to solve it.

The above approaches consider the current state of the decision making process of a committee where the committee is deciding whether to retain a vaccine strain from the previous flu season or select the most prevalent new strain circulating in the environment. In contrast, our approach is much more flexible in being able to identify emerging strains based on historical data and personalize the vaccine composition for a specific region/country. Moreover, none of these works quantify the efficacy of the vaccine in terms of morbidity and mortality rates as they do not prescribe vaccines for a particular geographical region.

1.2. Our contributions

This paper develops a holistic framework employing state-of-the-art methods in machine learning and optimization for prescribing influenza vaccine composition, and predicting the efficacy of the proposed vaccine in terms of morbidity and mortality rates for a particular region or a country. To the best of our knowledge, this is the first work that employs state-of-the-art methods in machine learning, namely tensor completion and ORTs, and robust optimization for the problem of optimizing influenza vaccine composition based on historical data.

The key contributions of this paper are summarized below.

1. We propose a novel tensor completion formulation that estimates a low-rank representation of the circulation rates of viruses across various regions around the world while restricting the low-rank component to not deviate much from its counterpart from a previous time period. Leveraging algorithms for tensor completion, we develop new algorithms to solve the corresponding restricted tensor completion problem.

2. We propose a set based uncertainty for the rates of circulation of viruses based on the low-rank component of the matrix factorization and formulate a robust prescriptive problem to choose vaccine composition that minimizes a worst-case weighted distance between the chosen vaccine-strains and viruses that are predicted to circulate in the future. We show that this problem can be reformulated as a second order cone optimization problem and show that it is both practically and theoretically tractable.

3. Through a retrospective study, we illustrate the effectiveness of our approach in terms of a weighted distance between the chosen vaccine composition and observed predominant circulating viruses during a flu season in comparison to vaccine compositions proposed by WHO for Northern Hemisphere, FDA for USA, EMA for Europe.

1.3. Notation

Lowercase and uppercase bold letters denote vectors and matrices, respectively. For a tensor $\mathbf{M} \in \mathbb{R}^{p\times q\times r}$, we denote the slices of the tensor as $\mathbf{M}^1, \mathbf{M}^2, \ldots, \mathbf{M}^r \in \mathbb{R}^{p\times q}$. A tensor *unfolding*, is an operation which essentially flattens the tensor into a matrix. The mode-1 unfolding of a tensor $\mathbf{M} \in \mathbb{R}^{p\times q\times r}$, denoted by $M_{(1)}$, is the $p \times qr$ matrix whose columns are the columns of $\mathbf{M}^1, \mathbf{M}^2, \ldots, \mathbf{M}^r$. Similarly the mode-2 unfolding, denoted by $M_{(2)}$, is the $q \times pr$ matrix whose columns are the transposed rows of $\mathbf{M}^1, \mathbf{M}^2, \ldots, \mathbf{M}^r$. The Frobenius norm of a matrix $\mathbf{U} \in \mathbb{R}^{m\times n}$ denoted by $\|\mathbf{U}\|_F$ is given as $\left(\sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}^2\right)^{1/2}$. We denote the set $\{1, 2, \ldots, n\}$ by $[n]$.

2. Data and Methods

In this section, we describe the data and methods used in our analysis to predict rates of circulation of predominant viruses in a flu season, and to predict morbidity and mortality rates given composition of a influenza vaccine and the observed rates of circulation of viruses in a flu season. In Section 2.1, we describe the data that we use in our analysis. In Section 2.2, we propose a novel tensor completion problem and present an algorithm

to model evolution of rates of circulation of viruses through different seasons. Finally, in Section 2.3, we review ORTs which we use to train regression trees for predicting morbidity and mortality rates using factors such as weighted distance between the vaccine strain and circulating viruses for each subtype of the influenza virus.

2.1. Data

Here, we describe the data that we use to inform influenza vaccine composition. The WHO has a network of laboratories around the world that contribute to the GISRS system which monitors and tracks properties of predominant circulating viruses using a test called Hemagglutinin Inhibition (HI) test. Below, we provide a short description of data compiled through GISRS, NIAID Influenza Research Database (IRD), The Centers for Disease Control and Prevention (CDC), and The European Centers for Disease Control and Prevention (ECDC).

1. **Rates of circulation of predominant viruses:** The historical rates of circulation of viruses are available in the form of a three-dimensional tensor (see Figure 1) $\mathbf{M} \in \mathbb{R}^{n \times m \times T}$, where n is the number of countries contributing to the WHO's GISRS system, m is the number of viruses in the data set and T is the number of flu seasons. Such a three-dimensional matrix is known as a three-dimensional *tensor*. Each entry in the tensor is given by

$$M^t_{cv} = \frac{N^t_{c,v}}{\sum_{v=1}^{m} N^t_{cv}}, \; c \in [n], v \in [m], t \in [T], \tag{1}$$

   where, N^t_{cv} is the number of observed cases of virus v found in tests performed in country c for flu season t. We denote the set of virus-strains belonging to influenza A (H1N1) as D_{H1N1}, and similarly, for influenza A (H3N2) and B as D_{H3N2} and D_{B} respectively. Each slice of the tensor representing a flu season from 1987 until 2018 ($T = 32$), and each slice consists of observed rates of circulation of $m = 1206$ viruses in about $n = 132$ countries.

2. **Estimates for morbidity and mortality rates due to influenza, and the corresponding flu vaccine compositions:** We compiled various estimates of morbidity and mortality rates for USA from the CDC website, for Europe from the ECDC website, and for Northern and Southern hemispheres from GISRS and Iuliano et al. (2018). The data for influenza vaccine composition over various years from 1987 until 2018 was obtained from the NIAID Influenza Research Database (IRD) (Bao et al., 2008).

3. **Antigenic properties of predominant circulating viruses:** At each of the laboratories that contribute to WHO's GISRS, circulating viruses during a season are subject to a HI (Hemagglutinin Inhibition) test which measures the ability of the antibodies (injected by a vaccine) to block the Hemagglutinin (HA) protein of the virus being tested from attacking healthy red blood cells. This data was obtained from IRD and ImmPort (Bhattacharya et al., 2018). For each virus u and vaccine-strain v, we have a corresponding HI value denoted by h_{uv}, where u belongs to either of

$D_{\text{H1N1}}, D_{\text{H3N2}}$, or D_{B}. Cai et al. (2012) proposed a distance metric between virus u and vaccine-strain v as follows,

$$d_{uv} = \log\left(\max_u (h_{uv})/h_{uv}\right), \tag{2}$$

where, $\max_u (h_{uv})$ is the maximum HI value for vaccine-strain v across all the viruses in the data set. Multiple measurements of virus-strain and vaccine-strain distances are available when antibodies raised against a viral strain are tested in multiple laboratories or at several time points, or when multiple antibodies are raised against the same strain. The resulting antigenic data set comprises of distances between $p = 1,377$ viruses and $q = 82$ vaccine strains.

2.2. Tensor completion

Here, we present our tensor completion formulation to model the evolution of rates of circulation of viruses across different geographies and flu seasons.

2.2.1. Formulation

First, we assume that a slice of the tensor $\mathbf{M}^t \in \mathbb{R}^{n\times m}, t \in \{1, 2, \ldots, 32\}$ can be expressed as the product $\mathbf{M}^t = \mathbf{U}\mathbf{S}^t\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n\times r}, \mathbf{V} \in \mathbb{R}^{m\times r}$ are latent spaces and $\mathbf{S}^t \in \mathbb{R}^{r\times r}$ is the low-rank component of the matrix factorization and r is assumed to be the rank of $\mathbf{M}^t$. Therefore, given the tensor of observed rates of circulating viruses across time and geographies, the tensor completion problem to estimate a low-rank matrix approximation for each time slice can be formulated as follows:

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V},\{\mathbf{S}_t\}} \sum_{t=1}^{T} \|\mathbf{M}^t - \mathbf{U}\mathbf{S}^t\mathbf{V}^\top\|_F \\ &\text{s.t.} \quad \|\mathbf{S}^t - \mathbf{S}^{(t-1)}\|_F \le \lambda,\ t \in \{2, 3, \ldots, T\}, \end{aligned} \tag{3}$$

where, $\|\cdot\|_F$ is the Frobenius norm and the parameter λ is chosen by the user to control for the deviation the low-rank component of the matrix factorization from its counterpart from a previous time period. Problem (3) is non-convex, therefore, we use an alternating optimization approach to solve it.

2.2.2. Tensor unfolding and alternating optimization

To estimate the latent spaces $\mathbf{U}$ and $\mathbf{V}$, we use the algorithm proposed in Bertsimas and Pawlowski (2019), which is based on Farias and Li (2019). In the first step, we construct the mode-1 unfolding $\mathbf{M}_{(1)} \in \mathbb{R}^{n\times mT}$, which which is a $n \times mT$ matrix whose columns are the columns of $\mathbf{M}^1, \mathbf{M}^2, \ldots, \mathbf{M}^T$ (the order of the columns does not matter). We then compute $\hat{\mathbf{U}}$ as the first r left singular vectors of $\mathbf{M}_{(1)}$. More precisely, assuming that $\mathbf{M}_{(1)}$ admits the singular value decomposition $\mathbf{M}_{(1)} = \mathbf{U}_1\Sigma_1\mathbf{V}_1^\top$, we set $\hat{\mathbf{U}}$ to be the columns of $\mathbf{U}_1$ corresponding to the r largest singular values. We denote this entire procedure with the shorthand

$$\hat{\mathbf{U}} = \text{SVD}(\mathbf{M}_{(1)}, r).$$

To estimate $\mathbf{V}$, we apply a similar procedure using the mode-2 unfolding $\mathbf{M}_{(2)}$, which is the $m \times nT$ matrix whose columns are the transposed rows of $\mathbf{M}^1, \mathbf{M}^2, \ldots, \mathbf{M}^T$. Therefore, the estimate of $\mathbf{V}$ is given by the r largest singular vectors of mode-2 unfolding $\mathbf{M}_{(2)}$ and we denote this procedure as follows

$$\hat{\mathbf{V}} = \text{SVD}(\mathbf{M}_{(2)}, r).$$

Given some estimates $\left(\hat{\mathbf{U}}, \hat{\mathbf{V}}\right)$, Problem (3) reduces to a second order cone optimization problem in $\{\mathbf{S}^t\}_{t=1}^T$ as follows:

$$\begin{aligned} \min_{\{\mathbf{S}_t\}} & \sum_{t=1}^{T} \|\mathbf{M}^t - \hat{\mathbf{U}}\mathbf{S}^t\hat{\mathbf{V}}^\top\|_F \\ \text{s.t.} \quad & \|\mathbf{S}^t - \mathbf{S}^{(t-1)}\|_F \leq \lambda, \ t \in \{2, 3, \ldots T\}. \end{aligned} \tag{4}$$

Problem (4) is a second order cone problem in Tr^2 variables and is tractable. Putting all of this together, we present Algorithm 1 to solve Problem (3) for a given value of rank r, parameter λ and maximum number of iterations K. We evaluate r and λ through k-fold cross validation with five folds.

2.3. Optimal Regression Trees

In order to predict the morbidity and mortality rates given a vaccine composition, we use a novel algorithm Optimal Regression Trees (ORTs) proposed in Bertsimas and Dunn (2017) to train predictive trees that combine state-of-the-art performance (at par with gradient boosted trees) and interpretability. Such tree structures are based on a few decision splits on variables of high importance, and can readily model non-linearities and interactions between variables.

To predict morbidity and mortality rates, we used the following predictors that quantify the ability of the vaccine-strains to hinder viruses' ability to attack healthy red blood cells,

1. Weighted distance between influenza A (H1N1) strain and the corresponding circulating viruses: $w_{\text{H1N1}} = \sum_{u \in D_{\text{H1N1}}} d_{uv} r_u$, where, v is the chosen vaccine-strain, r_u is the normalized rate of circulation of virus u among all predominant viruses in a particular flu season.

2. Weighted distance between influenza A (H3N2) strain and the corresponding circulating viruses: $w_{\text{H3N2}} = \sum_{u \in D_{\text{H3N2}}} d_{uv} r_u$.

3. Weighted distance between influenza B strain and the corresponding circulating viruses: $w_{\text{B}} = \sum_{u \in D_{\text{B}}} d_{uv} r_u$.

During the training process, we tuned the parameters to maximize performance on a separate holdout set to avoid overfitting.

In Figure 3, we present optimal regression trees that were trained on data from USA from 1988 until 2018 to predict morbidity and mortality rates using weighted distances between the vaccine-strains proposed by FDA and the actual viruses that circulated during

a particular flu season. The regression tree trained to predict morbidity has an accuracy (in terms of $\mathcal{R}^2$) of 0.77 and the one for mortality has an accuracy of 0.75. In both the trees, w_{H3N2} appears as an important variable which is expected as it is well known that influenza A (H3N2) strains highly volatile. Also, the positive coefficients of the weights w_{H1N1}, w_{H3N2} and w_{B} in the leaves of the regression trees signify that higher weighted distances between the vaccine-strains and predominant circulating viruses results in higher morbidity and mortality rates.

3. Optimizing the Vaccine Composition

In this section, we propose a set based uncertainty for the rates of circulation of viruses based on the low-rank component of the matrix factorization and formulate a robust prescriptive problem to choose vaccine composition that minimizes the worst-case weighted distance between the chosen vaccine-strain and the viruses that are predicted to circulate in the future. In Section 3.1, we describe a nominal formulation for choosing vaccine formulation a particular geographical location (prescribing country level, or region level influenza vaccine compositions) and in Section 3.2, we present the robust prescriptive counterpart of the nominal model.

3.1. Nominal model

Given estimates of antigenic distances d_{uv} (see equation (2)) between virus 'u' and vaccine-strain 'v', and let, $\mathbf{Y} = (y_{iu}) \in \mathbb{R}^{n\times q}$ denote the predicted rates of circulation of viruses, our goal is to choose a suitable vaccine-strain which has the smallest weighted distance with the predicted circulating viruses in a flu season. In order to select a suitable vaccine-strain to be included in the vaccine formulation for some location i, we propose to solve the following optimization problem:

$$\begin{aligned} \min_{z\in\{0,1\}^q} & \sum_{u=1}^{p}\sum_{v=1}^{q} z_v d_{uv} y_{iu} \\ \text{s.t.} \quad & \mathbf{Y} = \mathbf{U}\mathbf{S}_{T-1}\mathbf{V}^\top, \\ & \sum_{v=1}^{q} z_v = 1, \end{aligned} \tag{5}$$

where the term $\sum_{u=1}^{q} d_{uv} y_{iu}$ in the objective function represents a weighted distance between all viruses $u \in [p]$ predicted to be circulating with frequencies $\{y_{iu}\}_{u=1}^{p}$ and some vaccine-strain v.

3.2. Robust prescriptive model

Here, we present the robust prescriptive problem to choose vaccine composition for a particular geographical location. Given a low-rank decomposition of the Tensor $\mathbf{M}^t = \mathbf{U}\mathbf{S}^t\mathbf{V}^\top$, $t \in [T]$ containing rates of circulating viruses in different locations, we formulate a robust optimization problem to choose a vaccine strain that is robust to mutations during the time period the vaccines are manufactured and distributed.

Observe that Problem (5) is very sensitive to the predicted rates of circulation of viruses. Therefore, we propose using a set based uncertainty to capture any noise in the predictions from the tensor model. We define an uncertainty set for rates of circulation of viruses as follows,

$$\mathcal{U}_\lambda(\mathbf{U}, \hat{\mathbf{S}}, \mathbf{V}) = \left\{ \mathbf{Y} : \mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}^\top, \|\mathbf{S} - \hat{\mathbf{S}}\|_{F_2} \leq \lambda \right\}.$$

Therefore, instead of solving the nominal problem (5), we propose to solve the following robust optimization problem,

$$\begin{aligned} \mathcal{P}_i \quad = \quad & \min_{z \in \{0,1\}^q} \max_{\mathbf{Y} \in \mathcal{U}_\lambda(\mathbf{U}, \hat{\mathbf{S}}, \mathbf{V})} \sum_{u=1}^{p} \sum_{v=1}^{q} z_v d_{uv} y_{iu} \\ & \text{s.t.} \quad \sum_{v=1}^{q} z_v = 1. \end{aligned} \tag{6}$$

Lemma 1 (Neumann (1928)) *The min-max in Problem (6) can be interchanged, i.e.,*

$$\min_{\{\sum_{v=1}^q z_v = 1, \mathbf{z} \in \{0,1\}^q\}} \max_{\{\mathbf{Y} \in \mathcal{U}_\lambda(\mathbf{U}, \hat{\mathbf{S}}, \mathbf{V})\}} \sum_{u=1}^{p} \sum_{v=1}^{q} z_v d_{uv} y_{iu} = \max_{\{\mathbf{Y} \in \mathcal{U}_\lambda(\mathbf{U}, \hat{\mathbf{S}}, \mathbf{V})\}} \min_{\{\sum_{v=1}^q z_v = 1, \mathbf{z} \in \{0,1\}^q\}} \sum_{u=1}^{p} \sum_{v=1}^{q} z_v d_{uv} y_{iu}.$$

Using Lemma 1, observe that for a given geographical location represented by i, the objective function of Problem (6) is bilinear in $\mathbf{z}$ and Y_i, therefore, we can perform a "minimax swap" to obtain the following:

$$\begin{aligned} \mathcal{P}_i \quad = \quad & \max_{\mathbf{Y} \in \mathcal{U}_\lambda(\mathbf{U}, \hat{\mathbf{S}}, \mathbf{V})} \min_{z \in \{0,1\}^q} \sum_{u=1}^{p} \sum_{v=1}^{q} z_v d_{uv} y_{iu} \\ & \text{s.t.} \quad \sum_{v=1}^{q} z_v = 1, \end{aligned} \tag{7}$$

where the inner minimization problem can be reformulated as follows:

$$\begin{aligned} \mathcal{P}_i \quad = \quad & \max_{\epsilon, \mathbf{Y}, \mathbf{S}} \epsilon \\ \text{s.t} \quad & \epsilon \leq \sum_{u=1}^{p} d_{uv} y_{iu}, \ \forall v \in \{1, 2, \ldots, q\}, \\ & \mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}^\top, \\ & \|\mathbf{S} - \hat{\mathbf{S}}\|_{F_2} \leq \lambda. \end{aligned} \tag{8}$$

Problem (8) is a second order cone problem which can be further reduced by eliminating variables $\mathbf{Y}$, and can be solved using off-the-shelf solvers. The vaccine-strain prescribed by the model is given by ℓ, where $\ell = \arg\min_v \sum_{u=1}^{p} d_{uv} y^\star_{iu}$ and the worst-case weighted distance of the proposed vaccine-strains to the circulating viruses is given by $w_i^\star = \sum_{u=1}^{p} d_{u\ell} y^\star_{iu}$.

4. Results

In this section, we present the performance of vaccine compositions prescribed by the robust model in terms of predicted morbidity and mortality rates and compare it with that of WHO, FDA and EMA using Optimal Regression Trees. The morbidity and mortality rates are predicted using weighted distances between the vaccine strains and the observed rates of circulation viruses for each subtype of the influenza virus.

4.1. Morbidity and Mortality Rates

In Figure 3, we present optimal regression trees that were trained on data compiled from CDC, USA for flu seasons from 1988 until 2018 to predict morbidity and mortality rates using weighted distances between the vaccine-strains proposed by FDA and the actual virus strains that circulated during a particular flu season. For this analysis, we trained multiple regression trees with a moving window size of 20 years starting from 1988 and used these models to predict morbidity and mortality rates for the next flu season given some vaccine compositions along with observed rates of circulation of viruses.

In Tables 1, 2 and 3, we present a retrospective comparison of predicted mortality and morbidity rates based on the vaccine composition prescribed by WHO for Northern hemisphere, FDA for USA and EMA for Europe with our prescriptions for flu seasons during 2009-2018 respectively. We compare the effectiveness of prescribed vaccine composition by training regression trees using Optimal Regression Trees (ORTs) algorithm for predicting both morbidity and mortality using the following variables: (1) w_{H1N1}, a weighted distance between influenza A (H1N1) strain and the corresponding circulating viruses, and similar weighted distances for influenza A (H3N2) and influenza B viruses (2) w_{H1N1}, and (3) w_{B}. During the training process, we tuned the parameters to maximize performance on a separate holdout set to avoid overfitting.

Table 1: Retrospective 8-year comparison of number of illnesses and mortality under vaccine proposed WHO vs. robust prescriptive model for Northern hemisphere using Optimal Regression Trees (accuracy reported in terms of $\mathcal{R}^2$).

	Illnesses (in Millions)				Mortality (in Thousands)			
Season	Observed	WHO	Robust	Accuracy	Observed	WHO	Robust	Accuracy
2010-2011	3.2	3.27	**2.71**	0.68	427	432.4	**378.7**	0.74
2011-2012	2.6	3.52	**2.84**	0.68	362	435.1	**392.3**	0.73
2012-2013	4.2	3.37	3.37	0.71	436	434.3	434.3	0.76
2013-2014	3.9	3.73	**3.26**	0.69	438	481.7	**417.5**	0.78
2014-2015	3.7	3.58	**3.43**	0.72	451	466.8	**438.3**	0.72
2015-2016	3.6	3.34	**3.11**	0.74	463	434.1	**402.8**	0.77
2016-2017	4.4	3.92	**3.68**	0.71	482	504.2	**468.3**	0.70
2017-2018	4.2	4.56	**4.13**	0.67	461	483.6	**451.8**	0.73

In Table 1, we compare the effectiveness of vaccine compositions for Northern hemisphere against WHO. We observed that in seven of the eight flu seasons, vaccine prescribed by the robust model had a smaller number of morbidity and mortality cases. On average, vaccine compositions prescribed by the robust model could potentially lower morbidity by

Table 2: Retrospective 8-year comparison of number of severe illnesses and mortality under vaccine proposed FDA vs. robust prescriptive model for USA using Optimal Regression Trees (accuracy reported in terms of $\mathcal{R}^2$).

	Severe Illnesses (in Millions)				Mortality (in Thousands)			
Season	Observed	FDA	Robust	Accuracy	Observed	FDA	Robust	Accuracy
2010-2011	21	**24.4**	26.1	0.78	37	**36.7**	37.1	0.72
2011-2012	9.3	24.7	**20.8**	0.73	12	31.4	**29.6**	0.76
2012-2013	34	30.6	**28.2**	0.75	43	39.2	**36.7**	0.71
2013-2014	30	29.3	**26.4**	0.81	38	40.3	**38.4**	0.70
2014-2015	30	32.7	**28.5**	0.83	51	44.2	**27.4**	0.73
2015-2016	25	28.7	**25.9**	0.77	25	35.8	**33.1**	0.74
2016-2017	30	31.6	**28.8**	0.75	51	43.8	**41.6**	0.71
2017-2018	49	34.6	**31.7**	0.77	79	49.6	**46.4**	0.75

Table 3: Retrospective 8-year comparison of number of illnesses and mortality under vaccine proposed by EMA vs. robust prescriptive model for Europe using a Optimal Regression Trees (accuracy reported in terms of $\mathcal{R}^2$).

	Illnesses (in Millions)				Mortality (in Thousands)			
Season	Observed	EMA	Robust	Accuracy	Observed	EMA	Robust	Accuracy
2010-2011	12.1	13.1	**12.4**	0.78	82.5	82.8	**81.7**	0.68
2011-2012	8.6	11.7	**10.2**	0.76	68.6	79.4	**78.4**	0.71
2012-2013	14.3	14.5	**12.8**	0.79	87.5	85.2	**83.4**	0.67
2013-2014	13.7	**12.7**	13.4	0.76	79.3	**79.4**	81.2	0.69
2014-2015	13.4	13.6	**12.7**	0.75	103.4	87.2	**85.6**	0.74
2015-2016	11.6	13.3	**12.9**	0.76	76.3	85.8	**83.1**	0.71
2016-2017	12.4	13.2	**12.5**	0.72	105.3	92.8	**89.6**	0.67
2017-2018	15.8	14.2	**13.7**	0.74	132.7	97.6	**95.4**	0.64

11-14% and mortality by 8-11% over vaccine compositions proposed by WHO for Northern hemisphere.

Similarly, in Table 2, we compare the effectiveness of vaccine compositions for USA. Again, we observed that in seven of the eight flu seasons, vaccine compositions prescribed by the robust model had a smaller number of morbidity and mortality cases, and in only one season, we had greater morbidity and mortality cases. Vaccine compositions prescribed by the robust model could potentially lower morbidity by 8-10% and mortality by 6-9% over the ones proposed by FDA for USA.

Finally, in Table 3, we compare the effectiveness vaccine compositions for Europe. We observed that in seven of the eight flu seasons, vaccine compositions prescribed by the robust model had a smaller number of morbidity and mortality cases. Through ORTs, we show that vaccine compositions prescribed by the robust model could potentially lower morbidity by 10-12% and mortality by 9-11% over the ones proposed by EMA for Europe.

4.2. Prescribed Vaccine Formulations

Here, we compare the performance of the flu vaccine compositions prescribed by the robust model containing a type A (H1N1) virus in Section 4.2.1, a type A (H3N2) virus in Section 4.2.2, and a type B virus in Section 4.2.3 with the ones prescribed by WHO for Northern hemisphere, FDA for USA, and EMA for Europe in terms of weighted distance between the prescribed vaccine strains and the observed rates of circulation of viruses during the flu season.

4.2.1. Influenza Type A (H1N1) Virus

In Tables 4, 7 and 8 (in Appendix), we present a retrospective comparison of the H1N1-like virus strain as proposed by WHO for Northern hemisphere, FDA for USA and EMA for Europe with the ones prescribed by the robust model, respectively for the flu seasons during 2009-2018 in terms of a weighted distance w_{H1N1}.

In Table 4, we present vaccine strains prescribed by the robust model along with the ones proposed by WHO for Northern hemisphere. We observed that in four of the ten flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the observed circulating viruses and in four other seasons, the robust model prescribed the same vaccine strains as WHO.

In Table 7, we compare vaccine strains prescribed by the robust model with that of FDA for USA. We observe that in three of the ten flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the observed circulating viruses. And in five other flu seasons, the robust model and FDA proposed exactly same strains. Finally, in Table 8, we present vaccine strains prescribed by the robust model with that of EMA for Europe. We observe that in five of the ten flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the observed circulating viruses and in four other seasons, the robust model prescribed the same vaccine strains as EMA.

Table 4: Retrospective 10-year comparison of vaccine-strains (for influenza Type A (H1N1)-like virus) proposed by WHO vs. robust prescriptive model for Northern hemisphere.

	Weighted distance		Proposed vaccine	
Year	WHO	Robust	WHO	Robust
2009	3.51	**3.23**	A/Brisbane/59/2007	A/South Dakota/6/2007
2010	3.64	3.64	A/California/07/2009	A/California/07/2009
2011	3.38	3.38	A/California/07/2009	A/California/07/2009
2012	**3.19**	3.26	A/California/07/2009	A/Victoria/361/2011
2013	3.49	**3.29**	A/California/07/2009	A/Perth/56/2012
2014	3.03	3.03	A/California/07/2009	A/California/07/2009
2015	3.18	3.18	A/California/07/2009	A/California/07/2009
2016	3.67	**3.42**	A/California/07/2009	A/Hong Kong/4801/2014
2017	3.71	**3.44**	A/Michigan/45/2015	A/Wisconsin/67/2016
2018	**3.37**	3.51	A/Michigan/45/2015	A/Wisconsin/67/2016

4.2.2. Influenza Type A (H3N2) virus

In Tables 5, 9 and 10 (in Appendix), we present a retrospective comparison of the H3N2-like virus strain as proposed by WHO for Northern hemisphere, FDA for USA and EMA for Europe with the ones prescribed by the robust model, respectively for the flu seasons during 2009-2018 in terms of a weighted distance w_{H3N2}.

In Table 5, we compare vaccine strains prescribed by the robust model with that of WHO for Northern hemisphere. We observe that in six of the ten flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the observed circulating viruses and in two other seasons, the robust model prescribed the same vaccine strains as WHO.

In Table 9, we compare vaccine strains prescribed by the robust model with that of FDA for USA. We observed that in five of the ten flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the observed circulating viruses and in three other flu seasons, the robust model prescribed the same vaccine strains as FDA. Finally, in Table 10, we compare vaccine strains prescribed by the robust model with that of EMA for Europe. We observed that in six of the ten flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the observed circulating viruses and in two other seasons, the robust model prescribed exactly the same vaccine strains as EMA.

Table 5: Retrospective 10-year comparison of vaccine-strains (for influenza Type A (H3N2)-like virus) proposed by WHO vs. robust prescriptive model for Northern hemisphere.

	Weighted distance		Proposed vaccine	
Year	WHO	Robust	WHO	Robust
2009	3.59	**3.37**	A/Brisbane/10/2007	A/Uruguay/716/2007
2010	3.86	**3.27**	A/Perth/16/2009	A/California/7/2009
2011	3.41	3.41	A/Perth/16/2009	A/Perth/16/2009
2012	**3.32**	3.48	A/Victoria/361/2011	A/Victoria/210/2009
2013	3.65	**3.10**	A/Victoria/361/2011	A/Texas/50/2012
2014	3.42	**3.37**	A/Texas/50/2012	A/Wisconsin/15/2009
2015	3.15	**2.97**	A/Switzerland/9715293/2013	A/Norway/466/2014
2016	3.54	3.54	A/Hong Kong/4801/2014	A/Hong Kong/4801/2014
2017	4.08	**3.51**	A/Hong Kong/4801/2014	A/Singapore/INFIMH-16-0019/2016
2018	**3.36**	3.42	A/Singapore/INFIMH-16-0019/2016	A/Switzerland/8060/2017

4.2.3. Influenza Type B Virus

We present a retrospective comparison of the influenza B vaccine strain chosen by WHO, FDA and EMA with the ones prescribed by the robust model for the flu seasons during 2009-2018 in Table 6, 11 (in Appendix), respectively in terms of a weighted distance w_{B}.

In Table 6, we compare vaccine strains prescribed by the robust model with that of WHO for Northern hemisphere. We observed that in six of the ten flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the observed circulating viruses and in three other seasons, the robust model and WHO proposed exactly

same strains. In Table 11, we compare vaccine strains prescribed by the robust model with that of FDA for USA. We observed that in six of the ten flu seasons, vaccine strains prescribed by the robust model had a smaller weighted distance with the actual circulating viruses and in three other flu seasons, the robust model prescribed exactly same strains as FDA.

Table 6: Retrospective 10-year comparison of vaccine-strains (for influenza Type B virus) proposed by WHO vs. robust prescriptive model for Northern hemisphere.

	Weighted distance		Proposed vaccine	
Year	WHO	Robust	WHO	Robust
2009	3.21	3.21	B/Brisbane/60/2008	B/Brisbane/60/2008
2010	3.53	3.53	B/Brisbane/60/2008	B/Brisbane/60/2008
2011	3.34	**3.17**	B/Brisbane/60/2008	B/Wisconsin/01/2010
2012	3.21	3.21	B/Wisconsin/01/2010	B/Wisconsin/01/2010
2013	3.43	**3.22**	B/Massachusetts/02/2012	B/Massachusetts/02/2012
2014	3.31	**3.14**	B/Massachusetts/02/2012	B/Brisbane/33/2008
2015	3.24	**3.12**	B/Phuket/3073/2013	B/Brisbane/9/2014
2016	**3.23**	3.40	B/Brisbane/60/2008	B/Utah/09/2014
2017	3.54	**3.15**	B/Brisbane/60/2008	B/Phuket/3073/2013
2018	3.42	**3.16**	B/Phuket/3073/2013	B/Colorado/06/2017

5. Conclusions

In this paper, we have proposed a holistic framework employing state-of-the-art methods in machine learning and optimization to prescribe influenza vaccine composition based on historical data regarding circulating viruses in the population compiled through WHO's Global Influenza Surveillance and Response System (GISRS). Specifically, we proposed a novel tensor completion formulation that restricts low-rank component in the matrix factorization to not deviate from its counterpart from a previous time period. Using the estimates from tensor completion, we formulated a robust optimization problem to prescribe vaccine composition that is robust to rates of circulation of the viruses in region using a set based uncertainty on the low-rank component. Finally, we trained Optimal Regression Trees to predict both morbidity and mortality rates using weighted distances between vaccine viruses and circulating viruses during a flu season. Through various numerical experiments, we showed that our proposed vaccine compositions could potentially lower morbidity and mortality by 11-14%, 8-11% respectively over vaccine compositions proposed by WHO for Northern hemisphere, and lower morbidity and mortality by 8-10%, 6-9% over vaccine compositions proposed by FDA for USA, and finally, lower morbidity and mortality by 10-12%, 9-11% respectively over vaccine compositions proposed by EMA for Europe.

References

Joseph K Agor and Osman Y Özaltın. Models for predicting the evolution of influenza to inform vaccine strain selection. *Human Vaccines & Immunotherapeutics*, 14(3):678–683, 2018.

Yiming Bao, Pavel Bolotov, Dmitry Dernovoy, Boris Kiryutin, Leonid Zaslavsky, Tatiana Tatusova, Jim Ostell, and David Lipman. The influenza virus resource at the national center for biotechnology information. *Journal of Virology*, 82(2):596–601, 2008.

Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106(7): 1039–1082, 2017.

Dimitris Bertsimas and Colin Pawlowski. Tensor completion with noisy side information for the predcition of anti-cancer drug response. *Submitted*, 2019.

Sanchita Bhattacharya, Patrick Dunn, Cristel G Thomas, Barry Smith, Henry Schaefer, Jieming Chen, Zicheng Hu, Kelly A Zalocusky, Ravi D Shankar, Shai S Shen-Orr, et al. Immport, toward repurposing of open access immunological assay data for translational and clinical research. *Scientific data*, 5:180015, 2018.

Zhipeng Cai, Tong Zhang, and Xiu-Feng Wan. Antigenic distance measurements for seasonal influenza vaccine selection. *Vaccine*, 30(2):448–453, 2012.

Soo-Haeng Cho. The optimal composition of influenza vaccines subject to random production yields. *Manufacturing & Service Operations Management*, 12(2):256–277, 2010.

Vivek F Farias and Andrew A Li. Learning preferences with side information. *Management Science*, 2019.

Jiankui He and Michael W Deem. Low-dimensional clustering detects incipient dominant influenza strain clusters. *Protein Engineering, Design & Selection*, 23(12):935–946, 2010.

A Danielle Iuliano, Katherine M Roguski, Howard H Chang, David J Muscatello, Rakhee Palekar, Stefano Tempia, Cheryl Cohen, Jon Michael Gran, Dena Schanzer, Benjamin J Cowling, et al. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *The Lancet*, 391(10127):1285–1300, 2018.

Laura J Kornish and Ralph L Keeney. Repeated commit-or-defer decisions with a deadline: The influenza vaccine composition. *Operations Research*, 56(3):527–541, 2008.

Min-Shi Lee and Jack Si-En Chen. Predicting antigenic variants of influenza a/h3n2 viruses. *Emerging Infectious Diseases*, 10(8):1385, 2004.

Yu-Chieh Liao, Min-Shi Lee, Chin-Yu Ko, and Chao A Hsiung. Bioinformatics models for predicting antigenic variants of influenza a/h3n2 virus. *Bioinformatics*, 24(4):505–512, 2008.

Marta Luksza and Michael Lassig. A predictive fitness model for influenza. *Nature*, 507 (7490):57, 2014.

Richard A Neher and Trevor Bedford. nextflu: Real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*, 31(21):3546–3548, 2015.

Richard A Neher, Colin A Russell, and Boris I Shraiman. Predicting evolution from the shape of genealogical trees. *Elife*, 3:e03568, 2014.

J. von Neumann. The theory of social games. *Mathematical Annals*, 100(1):295–320, 1928.

Osman Y Özaltın, Oleg A Prokopyev, and Andrew J Schaefer. Optimal design of the seasonal influenza vaccine with manufacturing autonomy. *INFORMS Journal on Computing*, 30(2):371–387, 2018.

Alan S Perelson and George F Oster. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *Journal of theoretical biology*, 81(4):645–670, 1979.

L Steinbrück, TR Klingen, and AC McHardy. Computational prediction of vaccine strains for human influenza a (h3n2) viruses. *Journal of Virology*, 88(20):12123–12132, 2014.

Lars Steinbrück and Alice Carolyn McHardy. Inference of genotype–phenotype relationships in the antigenic evolution of human influenza a (h3n2) viruses. *PLoS Computational Biology*, 8(4):e1002492, 2012.

Ian A Wilson and Nancy J Cox. Structural basis of immune recognition of influenza virus hemagglutinin. *Annual review of Immunology*, 8(1):737–787, 1990.

Joseph T Wu, Lawrence M Wein, and Alan S Perelson. Optimization of influenza vaccine selection. *Operations Research*, 53(3):456–476, 2005.

Appendix A.

A.1. Tensor: Structure and Unfoldings

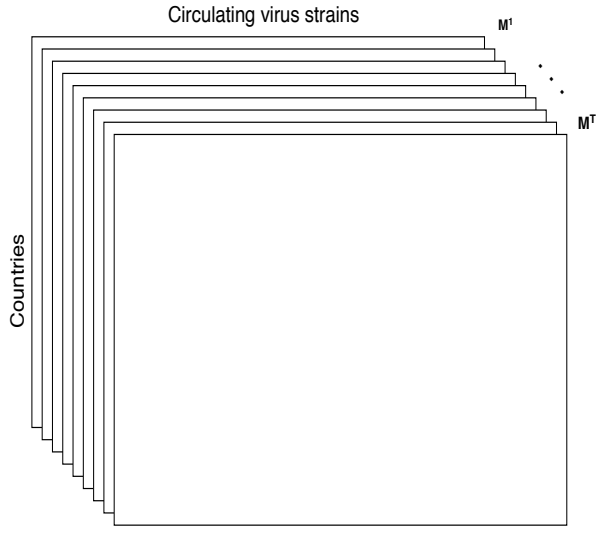


Figure 1: Rates of circulation of predominant viruses across different times represented as slices of a tensor $\mathbf{M}$.

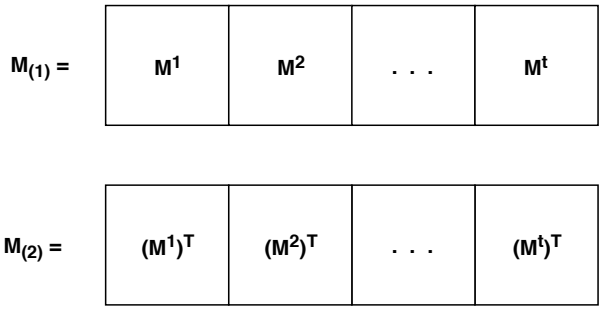
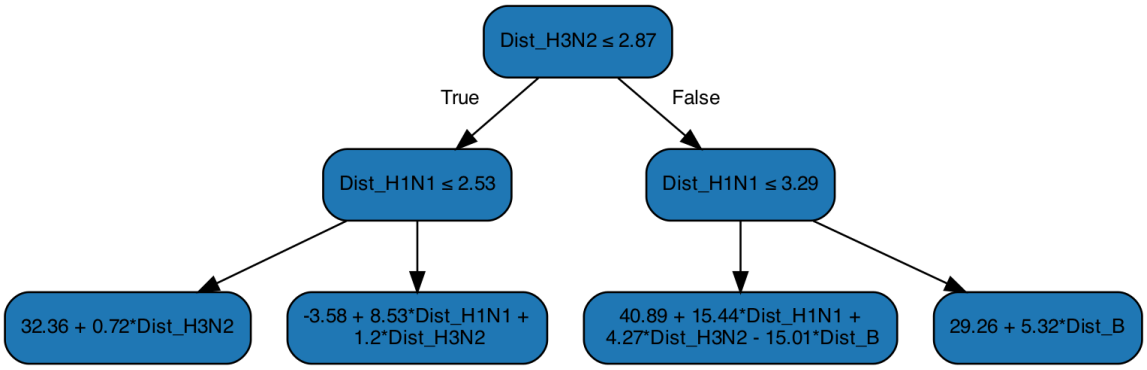


Figure 2: Example of mode-1 and mode-2 unfoldings of a tensor $\mathbf{M}$.

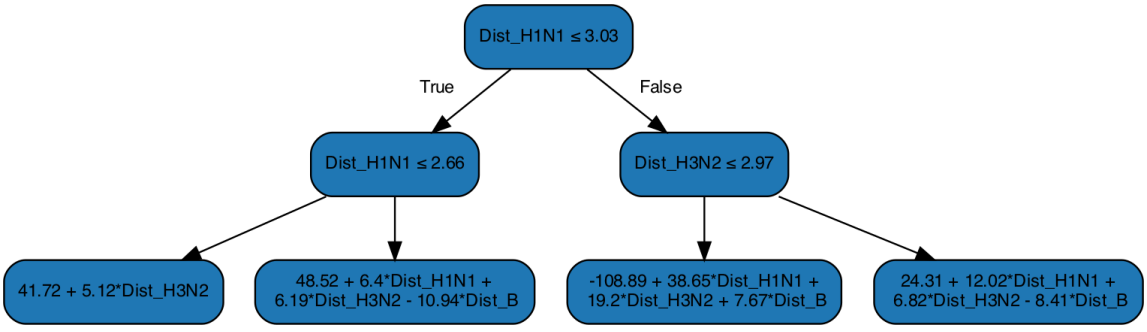
A.2. Optimal Regression Trees

Figure 3: Optimal regression trees to predict morbidity (in Millions) and mortality (in Thousands) in USA for flu seasons from 1988 till 2018. The variables $\text{Dist}_{\text{H1N1}}$, $\text{Dist}_{\text{H3N2}}$ and Dist_{B} denote weighted distances between the predominant circulating viruses and influenza A (H1N1), A (H3N2) and B strains, respectively.



(*a*)

Figure 4: Regression tree for predicting morbidity (accuracy in terms of $\mathcal{R}^2$: 0.77).



(*a*)

Figure 5: Regression tree for predicting mortality (accuracy in terms of $\mathcal{R}^2$: 0.75).

A.3. Tensor completion

Algorithm 1 Tensor completion (Problem (3))

Data: Incomplete tensor $\mathbf{X} \in \mathbb{R}^{m \times n \times T}$, Rank r, parameter λ and max iterations K.
Result: $\left(\hat{\mathbf{U}}, \{\hat{\mathbf{S}}^t\}_{t=1}^T, \hat{\mathbf{V}}\right)$
$\hat{\mathbf{Y}}_0 \leftarrow \mathbf{X}$ and $k \leftarrow 1$.
while $k < K$ **do**
  $\hat{\mathbf{U}}_k = \text{SVD}\left(\hat{\mathbf{Y}}_{k-1,(1)}, r\right)$
  $\hat{\mathbf{V}}_k = \text{SVD}\left(\hat{\mathbf{Y}}_{k-1,(2)}, r\right)$
  $\{\hat{\mathbf{S}}_k\}_{t=1}^T =$ Solve Problem (4) with estimates $(\hat{\mathbf{U}}_k, \hat{\mathbf{V}}_k)$ and parameter λ
  $\hat{\mathbf{Y}}_k^t = \hat{\mathbf{U}}_k \hat{\mathbf{S}}_k^t \hat{\mathbf{V}}_k^T,\ t \in [T]$
  $k = k + 1$
end
return $\left(\hat{\mathbf{U}}_K, \{\hat{\mathbf{S}_K^t}\}_{t=1}^T, \hat{\mathbf{V}}_K\right)$.

A.3. Prescribed Vaccine Compositions

Table 7: Retrospective 10-year comparison of vaccine-strains (for H1N1-like virus) proposed by FDA vs. robust prescriptive model for USA.

	Weighted distance		Proposed vaccine	
Year	FDA	Robust	FDA	Robust
2009	**3.22**	3.35	A/Brisbane/59/2007	A/Wisconsin/67/2005
2010	3.34	3.34	A/California/07/2009	A/California/07/2009
2011	3.09	3.09	A/California/07/2009	A/California/07/2009
2012	3.27	**3.14**	A/California/07/2009	A/South Australia/55/2014
2013	3.59	**3.37**	A/California/07/2009	A/Wyoming/03/2010
2014	3.48	3.48	A/California/07/2009	A/California/07/2009
2015	3.62	3.62	A/California/07/2009	A/California/07/2009
2016	3.30	**3.06**	A/California/07/2009	A/Brisbane/10/2012
2017	**3.17**	3.53	A/Michigan/45/2015	A/Brisbane/10/2007
2018	3.26	3.26	A/Michigan/45/2015	A/Michigan/45/2015

Table 8: Retrospective 10-year comparison of vaccine-strains (for H1N1-like virus) proposed by EMA vs. robust prescriptive model for Europe.

	Weighted distance		Proposed vaccine	
Year	EMA	Our model	EMA	Our model
2009	3.51	3.23	A/Brisbane/59/2007	A/Brisbane/59/2007
2010	3.64	3.64	A/California/07/2009	A/California/07/2009
2011	3.38	3.38	A/California/07/2009	A/California/07/2009
2012	3.19	**3.26**	A/California/07/2009	A/Victoria/361/2011
2013	3.49	**3.29**	A/California/07/2009	A/Victoria/361/2011
2014	3.03	**3.03**	A/California/07/2009	A/Christchurch/16/2010
2015	3.18	3.18	A/California/07/2009	A/California/07/2009
2016	3.67	**3.42**	A/California/07/2009	A/Michigan/45/2015
2017	3.71	**3.44**	A/Michigan/45/2015	A/Brisbane/10/2012
2018	**3.37**	3.51	A/Michigan/45/2015	A/Wisconsin/67/2016

Table 9: Retrospective 10-year comparison of vaccine-strains (for H3N2-like virus) proposed by FDA vs. robust prescriptive model for USA.

	Weighted distance		Proposed vaccine	
Year	FDA	Robust	FDA	Robust
2009	3.48	3.48	A/Brisbane/10/2007	A/Brisbane/10/2007
2010	3.64	**3.36**	A/Perth/16/2009	A/California/7/2009
2011	3.34	**3.07**	A/Perth/16/2009	A/Uruguay/716/2007
2012	**3.09**	3.32	A/Victoria/361/2011	A/Wisconsin/15/2009
2013	3.75	**3.57**	A/Victoria/361/2011	A/Texas/50/2012
2014	3.46	3.46	A/California/7/2009	A/California/7/2009
2015	3.23	**3.12**	A/Switzerland/9715293/2013	A/Norway/466/2014
2016	3.21	**3.06**	A/Hong Kong/4801/2014	A/Stockholm/6/2014
2017	2.92	2.92	A/Singapore/INFIMH-16-0019/2016	A/Singapore/INFIMH-16-0019/2016
2018	3.04	3.04	A/Singapore/INFIMH-16-0019/2016	A/Singapore/INFIMH-16-0019/2016

Table 10: Retrospective 10-year comparison of vaccine-strains (for H3N2-like virus) proposed by EMA vs. robust prescriptive model for Europe.

	Weighted distance		Proposed vaccine	
Year	EMA	Robust	EMA	Robust
2009	**3.22**	3.37	A/Brisbane/10/2007	A/Brisbane/60/2008
2010	3.46	**3.29**	A/Perth/16/2009	A/California/7/2009
2011	3.53	3.53	A/Perth/16/2009	A/Perth/16/2009
2012	3.21	**3.06**	A/Massachusetts/2/2012	A/Victoria/210/2009
2013	3.67	**3.39**	A/Victoria/361/2011	A/Wisconsin/1/2010
2014	3.35	**3.03**	A/Texas/50/2012	A/Wisconsin/15/2009
2015	3.42	**3.17**	A/Switzerland/9715293/2013	A/Phuket/3073/2013
2016	3.43	3.43	A/Hong Kong/4801/2014	A/Hong Kong/4801/2014
2017	**3.38**	3.51	A/Singapore/INFIMH-16-0019/2016	A/Michigan/45/2015
2018	3.35	**3.12**	A/Singapore/INFIMH-16-0019/2016	A/Switzerland/8060/2017

Table 11: Retrospective 10-year comparison of vaccine-strains (for influenza Type B virus) proposed by FDA vs. robust prescriptive model for USA.

	Weighted distance		Proposed vaccine	
Year	FDA	Robust	FDA	Robust
2009	3.25	3.25	B/Brisbane/60/2008	B/Brisbane/60/2008
2010	3.64	3.64	B/Brisbane/60/2008	B/Brisbane/60/2008
2011	3.38	**3.21**	B/Brisbane/60/2008	B/Wisconsin/01/2010
2012	3.17	3.17	B/Wisconsin/01/2010	B/Wisconsin/01/2010
2013	3.49	**3.26**	B/Massachusetts/02/2012	B/Malaysia/2506/2009
2014	3.12	**2.97**	B/Massachusetts/02/2012	B/Brisbane/33/2008
2015	3.18	**3.04**	B/Phuket/3073/2013	B/Brisbane/9/2014
2016	3.67	**3.42**	B/Brisbane/60/2008	B/Utah/09/2014
2017	3.71	**3.44**	B/Brisbane/60/2008	B/Phuket/3073/2013
2018	**3.37**	3.51	B/Phuket/3073/2013	B/Colorado/06/2017