IOWA STATE UNIVERSITY

Department of Statistics

Using Mixture Distributions in Collaborative Forecasting of COVID-19

Spencer Wadsworth & Jarad Niemi

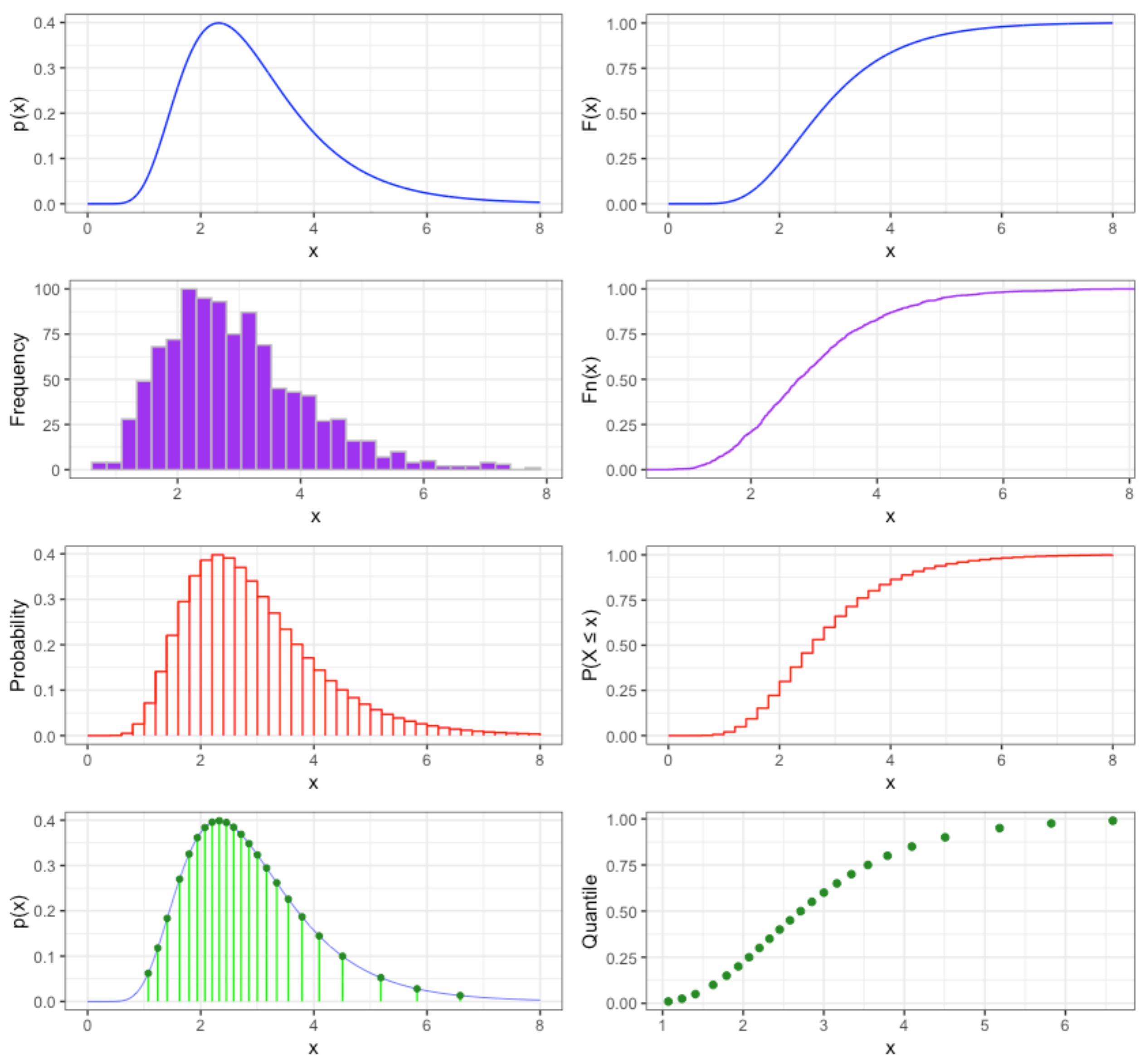
Introduction

The COVID-19 Forecast Hub was established in 2020 as a collaborative effort to forecast the coronavirus pandemic. Dozens of teams have been involved, each submitting weekly probabilistic forecasts of the spread of the virus. From the weekly forecasts, a single ensemble forecast is constructed and shared with the CDC for use in public reports. In order to assess the various forecasts against one another and to construct from them an ensemble forecast, each submitted forecast must share a common format. Each forecast is submitted as a set of 23 *quantiles* and their corresponding values.

Other collaborative forecast projects use different formats for representation such as *binned probabilities* used in the CDC's annual flu forecasting competition. Other formats include *parametric distributions* and *sample distributions.*

We propose that collaborative forecast projects utilize a *discrete mixture distribution* format as a preferable alternative to the four mentioned above.

Forecast representations used in disease outbreaks

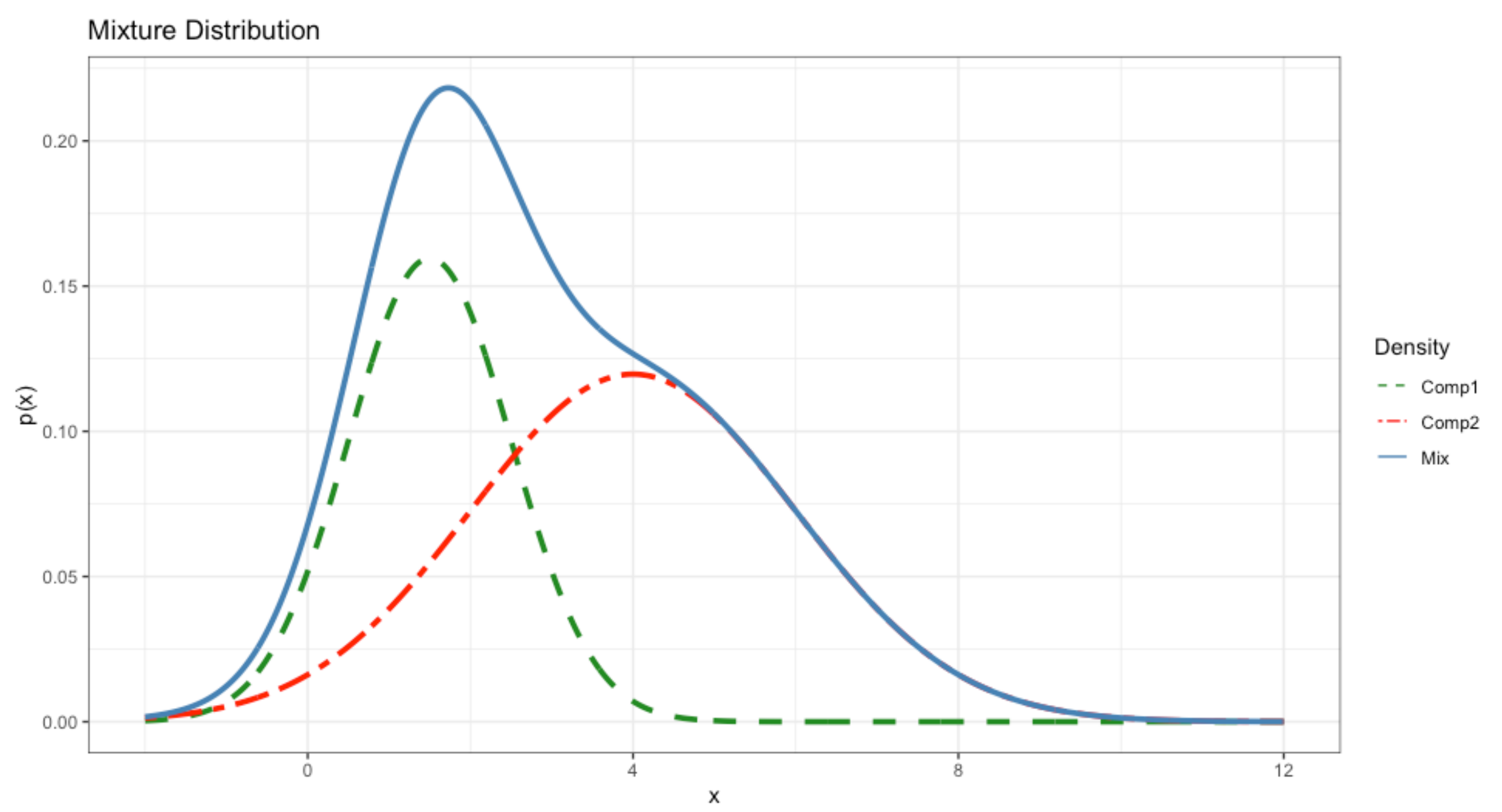


This image illustrates four possible formats used in collaborative forecasting of disease outbreaks. The left column illustrates density functions for a parametric distribution, sample distribution, binned distribution and quantiles respectively. The right column illustrates the corresponding probability function.

Mixture distribution construction

$$p^M(x) = \sum_{c=1}^{C} w_c p_c(x)$$

The density function of a continuous mixture distribution is the weighted sum of density functions of parametric distributions. Such distributions allow for many distributional shapes while requiring a relatively small amount of information to define the full distribution.

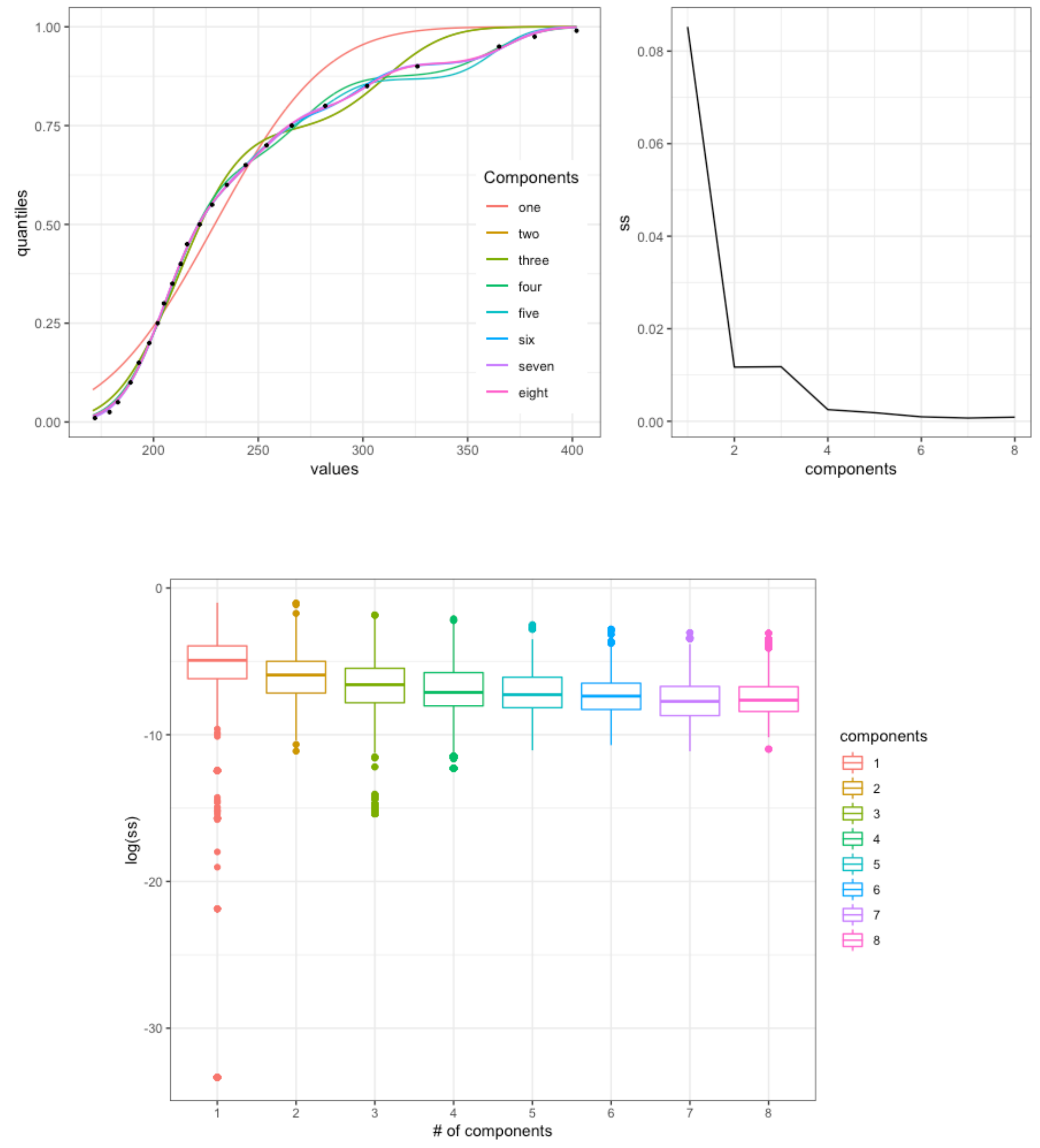


Representation comparison

	Scoring	Flexibility/information	Ensemble	Storage
Parametric distribution	LogS, CRPS, IS	Limited to common distribution families. Infinite resolution	Model averaging	Low 3-6 values per forecast
Sample distribution	CRPS, IS, and LogS after smoothing	Any shape. Resolution depends on sample size	Model averaging (after smoothing)	Hundreds-thousands of values per forecast
Bin distribution	LogS, CRPS, IS	Any shape allowed but limited by binning scheme	Model averaging	Depends on binning scheme. Dozens to hundreds of values per forecast
Quantiles	IS, WIS	Shape unknown but may still provide decent information. No tail information	Quantile averaging	Depends on requested quantiles. Maybe dozens of values per forecast
Discrete mixture distribution	LogS, CRPS, IS	With sufficient number of components, may approximate any distribution shape. Infinite resolution	Model averaging	Depends on number of components permitted/used. Few to dozen values per forecast

This table shows how discrete mixture distributions compare with the other four representations mentioned in terms of proper scoring, information contained in a forecast, ensemble construction, and computer storage. The discrete mixture distribution may be scored using the most commonly used proper scoring rules, has high flexibility in terms of shape, contains all distributional information using a relatively small amount of computer storage, and ensemble distributions may be constructed using the most common methods.

Can a quantile forecast be approximated by a mixture of normal distributions?



These two figures show the results of fitting mixture of normal distributions from 1 to 8 components to two different forecasts submitted to the COVID-19 forecast hub. The plot on top shows the results of fitting the 1 to 8 component mixture distributions to a single COVID-19 forecast. The fit clearly improves as the number of components increases. The boxplots below are of the log sum of squares of fitting 1 to 8 component mixture distributions to

Conclusions

Using a mixture distribution forecast format in collaborative forecast projects may be a preferable alternative to other commonly used representations. There is also evidence that quantile forecasts used in the COVID-19 Forecast Hub may be closely approximated by mixture distributions with an increasing number of components. However, a transition from a currently used format such as quantiles to mixture distributions may be difficult to make. Further research on the benefits of mixture distributions and on the development of tools for evaluating them is recommended.

References: Bracher, J., Ray, E. L., Gneiting, T., and Reich, N. G. (2021). Evaluating epidemic forecasts in an interval format. PLoS computational biology, 17(2):e1008618.

COVID-19 Forecast Hub: https://covid19forecasthub.org/