Non-parametric Bayesian Causal Modeling of the SARS-CoV-2 Viral Load Distribution vs. Patient's Age

A Preprint

Matteo Guardiani [1,2,*], Philipp Frank [1,2], Andrija Kostić [1,2],
Gordian Edenhofer [1,2], Jakob Roth [1,2], Berit Uhlmann [4], and Torsten Enßlin [1,2,3]

[1] Max Planck Institute for Astrophysics, Karl-Schwarzschild-Straße 1, 85748 Garching, Germany
[2] Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539 Munich, Germany
[3] Excellence cluster ORIGINS, Boltzmannstraße 2, 85748 Garching, Germany
[4] Süddeutsche Zeitung, Hultschiner Straße 8, 81677 Munich, Germany

May 31, 2021

Abstract

The viral load of patients infected with SARS-CoV-2 varies on logarithmic scales and possibly with age. Controversial claims have been made in the literature regarding whether the viral load distribution actually depends on the age of the patients. Such a dependence would have implications for the COVID-19 spreading mechanism, the age-dependent immune system reaction, and thus for policymaking. We hereby develop a method to analyze viral-load distribution data as a function of the patients' age within a flexible, non-parametric, hierarchical, Bayesian, and causal model. This method can be applied to other contexts as well, and for this purpose, it is made freely available. The developed reconstruction method also allows testing for bias in the data. This could be due to, e.g., bias in patient-testing and data collection or systematic errors in the measurement of the viral load. We perform these tests by calculating the Bayesian evidence for each implied possible causal direction. When applying these tests to publicly available age and SARS-CoV-2 viral load data, we find a statistically significant increase in the viral load with age, but only for one of the two analyzed datasets. If we consider this dataset, and based on the current understanding of viral load's impact on patients' infectivity, we expect a non-negligible difference in the infectivity of different age groups. This difference is nonetheless too small to justify considering any age group as noninfectious.

Keywords SARS-CoV-2 · COVID-19 · Viral load · Age · Causal inference · Desity estimation

1 Introduction

Children do not seem to be major drivers in the transmission of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in the general population [1]. However, the exact degree to which children and adolescents get infected by, and are able to transmit the virus is not yet well known. Their role in the community spread depends on their susceptibility, symptoms, viral load, social contact patterns, behavior, and existing mitigation strategies as schools and daycares closings. Among all these variables, the viral load plays a fundamental role. The viral load might help to predict disease severity [2] and mortality [3–5] and can serve as a proxy for the infectivity of the patient [6–8]. The severity of a disease, its infectivity, and its mortality are certainly fundamental parameters that must be considered when deciding on best-practice preventative measures to fight the pandemic spread. Research in this direction can enable truly data-driven policymaking like, for example, school openings and focused lockdowns. For this scope, it is important to understand how the viral load depends on the patients' age.

In this work, we examine viral load as a proxy for infectivity. We reanalyze the age-stratified viral load data from Jones et al. [9] in order to better understand the actual relation between these variables. We do this in the hope of gaining

[*] Correspondence to: matteani@mpa-garching.mpg.de

insight into fundamental differences reported in the literature regarding the relationship between viral load and age. We achieve this goal by developing a flexible, non-parametric, causal, and Bayesian model to reconstruct the conditional probability density function (PDF) of the viral load given the patient's age. The developed method is a second central result of this work: it can be applied to future studies on SARS-CoV-2, to similar data from other diseases, and also to many causally connected quantities in very different contexts.

The non-parametric PDF reconstruction is regularized by mild assumptions on the smoothness of the underlying statistical processes. In particular, we assume that the log-densities are Gaussian-process realizations, drawn with an a priori flexible correlation kernel parametrized by a Matérn family correlation function. The parameters of this correlation function are then inferred along with the PDF through a variational inference algorithm. To achieve this, we adapt methods developed for information field theory, the information theory for fields [10, 11]. In this context, fields are understood as spatially varying (physical) quantities. The reconstructed PDFs are regarded as scalar fields whose values are defined at each point of the two-dimensional space spanned by age and viral load and represent the probability of observing a given age and viral load. The toolkit of information field theory has proven itself to be successful in a wide range of applications, ranging from 3D tomography [12], over time-resolved astronomical imaging [13], to causality inference [14].

Outline The rest of this work is structured as follows: in Sec. 2, we discuss the state of the art of the research on the assessment of the viral load-vs-age dependence. In Sec. 3, we describe the model, how the causal structure is introduced and built into it, and the inference scheme we adopt. In Sec. 4, we describe how the data has been acquired and processed. We then outline the main results of our study in Sec. 5, focusing on their impact on the infectivity of SARS-CoV-2. Finally, in Sec. 6 we highlight potential limitations and identify possible future work directions.

2 Related work

Since the outbreak of SARS-CoV-2, efforts have been made in order to understand whether certain age groups are more susceptible than others. This could either mean that people from such age groups are more likely to get infected or that they show more severe symptoms compared to older or younger individuals. In addition to this, patients from specific age groups could be more infectious than others, hence more likely vehiculating the disease. To shed light on these problems, viral load – which is a proxy for infectivity – can be a useful tool. It is measured by reverse transcription PCR (RT-qPCR) assays from nasopharyngeal and oropharyngeal swabs via the so-called cycle threshold (Ct) value The viral load is the virus concentration in the upper respiratory tract and it is usually expressed as the number of viral RNA copies per mL of sample or entire swab specimen or simply by the Ct value.

Several works have analyzed whether viral loads differ between children and adults [1, 15–20], and between young children and adolescents [21–23], and have led to conflicting results. At least four studies from different countries have concluded that SARS-CoV-2 viral RNA loads among children and adults were comparable [1, 15, 16, 19]. In these studies, dependence between age and viral load has been tested using one and two-way analysis of variance (ANOVA) [15] and median of the viral load. A further study [17] found that mean and median viral load values did not vary conspicuously by age but noted that the highest values were measured in patients born from 1995 to 2009. In contrast, at least two studies reported significant differences in viral loads of young children and adults [18–20]. Euser et al. [18] suggested an approximately 16-fold higher viral load in the oldest age group (> 79 years), compared to the youngest age group (< 12 years). Here, age-group differences in the viral load distribution are assessed making use of the Kruskal-Wallis test and linear regression. Another work [20] estimated the amount of SARS-CoV-2 in the upper respiratory tract of young children (< 5 years) to be 10-fold to 100-fold greater than in adults whereas a work from Zachariah et al. [23] showed that mean viral load was significantly higher in infants (< 1 year) as compared to older children and adolescents.

Finally, one of the largest and most widely followed studies on the subject – even though the number of children and adolescents included is fairly small - was carried out by Jones et al. [9] in early 2020. Dependence between viral load and age has been tested for different age groups both as categorical data and treating age as a continuous variable. In order to compare the viral load of different age categories, the categorical data has been analyzed in a parametric (Welch's T-test), non-parametric (Mann-Whitney rank test), and Bayesian fashion (modeling viral loads as a mixture of gamma distributions). When considering age as a continuous variable, viral loads have been predicted from age, type of PCR system, and age-PCR system interaction. This study from Jones et al. [9] did not reveal large differences in the viral loads of different age groups, a result that was publicly debated in Germany for its possible implications for school opening policies. In the following, we reanalyze this data. In a very recent publication, Jones et al. [24] extended their initial version of the study. In this newer version, they make use of thin-plate spline regression to conclude that children and adolescents have a slightly lower viral load than adults, but that this difference is unlikely to be clinically relevant.

3 Model design

In Sec. 2 we described the state of the art of statistical analyses performed in order to investigate the age dependence of the viral load. These analyses mostly rely on variance tests or correlation assessments. In the following, we motivate the need for a causal model of the viral load distribution. In fact, in order to explain the age and viral load data collected by Jones et al. [9], a well-behaved model should incorporate basic knowledge about the causal relation between age and viral load. It should furthermore allow questioning whether the viral load distribution depends on the patients' age and quantify the strength of such dependence, if it exists. We consequently develop a non-parametric causal model and apply it to data.

3.1 Causal structure

The analyzed dataset $d = \{(i, x_i, y_i)\}_{i=1}^{N}$ consists of indexed pairs of age $x_i = \text{age}_i/\text{years}$ and log-viral load $y_i = \log_{10}(\text{viral RNA copies}_i/\text{ml})$ values for each of the N infected patients, extracted from Fig. 6 in [9]. The dataset is shown in Fig. 3. For simplicity, we refer to y as the "viral load", where 10^y are the number of viral RNA copies per milliliter of sample or entire swab specimen. We assume that the data points are Poisson process counts drawn from an underlying stationary density distribution $\varrho(x, y)$. Our final reconstruction of the density $\varrho(x, y)$ is also shown in Fig. 3. To model this density, we express it in terms of the underlying age distribution $\varrho(x)$ of the patients times the conditional PDF $p(y|x)$ of the viral load y given the patient's age x,

$$\varrho(x, y) = \varrho(x)\; p(y|x). \tag{1}$$

In Eq. 1 the causal direction $x \to y$ (age influences viral load) is implicitly introduced. Even though $x \to y$ might appear to be the most intuitive causal direction, since the immune system reaction depends on age and thus age should affect the viral load, we note that selection effects could introduce different apparent causal structures onto the data. For example, if the data would have been collected in such a way that the viral load (y) was the deciding factor for whether a patient would enter the data sample, with an age (x) dependent threshold, the viral load would impact the age distribution in the sample, leading to an apparent $y \to x$ causal structure. As such selection effects cannot be fully excluded for the analyzed data (see discussion in [9]), we will calculate the Bayesian evidence for the possible causal relations $x \to y$, $y \to x$, and $x \perp y$ (*i.e.* x and y are independent, $p(y|x) = p(y)$).

To model the $x \to y$ causal direction, we need to model the age distribution $\varrho(x)$ of the infected patients according to the causal structure introduced in Eq. 1. Lacking knowledge on the exact details of the patient selection process, we assume $\varrho(x)$ to be a log-normally distributed random variable. The log-normal distribution is a natural choice since an age density is by definition a strictly positive and continuous quantity. Another natural assumption is the absence of abrupt changes, since no sharp age-selecting processes are expected to have shaped it. We fulfill these assumptions with the choice

$$\varrho(x) = \varrho_0 e^{f(x)}, \tag{2}$$

where $\varrho_0 = N/100$ is a reference density and $f : [0, 100] \mapsto \mathbb{R}$ a smooth function centered around zero. We accordingly assume f to be drawn from a zero centered Gaussian process with a prior covariance F

$$\mathcal{P}(f) = \mathcal{G}(f, F) := \frac{1}{\sqrt{2\pi F}} \exp\left(-\frac{1}{2} f^\dagger F^{-1} f\right). \tag{3}$$

The covariance

$$F_{xx'} = \langle f(x)\, f(x') \rangle_{(f)} := \int \mathcal{D}f\; \mathcal{P}(f)\, f(x)\, f(x') \tag{4}$$

determines the degree of smoothness of the logarithmic distribution function, as well as the characteristic length scale and the amplitude of its variations. We assume this correlation structure to be translation invariant $F_{xx'} = F(x - x')$, since we only expect it to depend on age differences - and not on a particular age value - and parametrize it with a Matérn kernel. Invoking the Wiener-Khinchin theorem, we can represent such a translation invariant correlation function in Fourier space with a spectral density of

$$P_f(k) = \frac{a_f^2}{\left[1 + (k/k_f)^2\right]^{\gamma_f/2}}, \tag{5}$$

with a_f specifying the amplitude of the variations in f, $1/k_f$ the characteristic length-scale above which the variations become uncorrelated, and γ_f the spectral index, which determines the smoothness of the variations. We infer all three covariance parameters $p_f := (a_f, k_f, \gamma_f)$ from the data. In order to ensure that the model is flexible enough to fit the data, we set mildly informative priors on the covariance parameters.[2] We denote by $\mathcal{P}(f|p_f)$ the probability of a specific realization of f given the Matérn kernel parameters p_f, as described by Eqs. 3 to 5.

Next, we have to specify the distribution of the viral load given the age, $p(y|x)$. We do this in such a way that independent distributions, for which $p(y|x) = p(y)$, are singled out, whereas dependent ones, for which $p(y|x) \neq p(y)$, are only introduced if strictly required by the data. To this end, we choose

$$p(y|x) \propto \frac{e^{g(y)+h(x,y)}}{\int e^{h(\tilde{x},y)}\,\mathrm{d}\tilde{x}}. \tag{6}$$

The function $g(y)$ describes structures that depend only on the viral load y, whereas $h(x, y)$ models any entanglement between the age x and the viral load y.

In principle, the function $h(x, y)$ can represent any function $p(y|x)$ without the necessity of any structure represented via $g(y)$. To ensure that $g(y)$ captures all the strictly y-dependent structures it can represent, we prevent $h(x, y)$ from encoding any y-only structure. This is done in Eq. 6 with the denominator$\int e^{h(\tilde{x},y)}\,\mathrm{d}\tilde{x}$ that removes from $p(y|x)$ any y-only structure contained in $h(x, y)$.

To verify that $h(x, y)$ has indeed this desired property - i.e. that it cannot encode any structure that could be represented by $g(y)$ - it is simply possible to substitute $h(x, y) \mapsto h'(x, y) := h(x, y) + g'(y)$ in Eq. 6, where $g'(y)$ depends only on y. This results in h and h' leading to the same conditional PDF

$$\begin{aligned} p'(y|x) &\propto \frac{e^{g(y)+h'(x,y)}}{\int e^{h'(\tilde{x},y)}\,\mathrm{d}\tilde{x}} = \frac{e^{g(y)+h(x,y)+\cancel{g'(y)}}}{\cancel{e^{g'(y)}}\int e^{h(\tilde{x},y)}\,\mathrm{d}\tilde{x}} \\ &= \frac{e^{g(y)+h(x,y)}}{\int e^{h(\tilde{x},y)}\,\mathrm{d}\tilde{x}} \propto p(y|x) \end{aligned} \tag{7}$$

and therefore we conclude that only $g(y)$ models y-only dependent features.

For vanishing h, we get

$$p(y|x)\bigg|_{h(x,y)=0} \propto e^{g(y)} \propto p(y), \tag{8}$$

which implies independence, $y \perp x$. Thus, a non-trivial $h(x, y)$ models the causal influence $x \rightarrow y$ and causal independence is represented by $h(x, y) \equiv 0$.

The distinction between g and h would be meaningless without a prior choice that favors independence between x and y. Hence, we assume g and h to be drawn from zero centered Gaussian processes. In this way, without any information coming from the data, the most likely realizations of these functions are the zero functions $g(y) = h(x, y) \equiv 0$. However, if any structure in the data requests structure in the marginal distribution $p(y)$, such structure can only be represented by a non-zero $g(y)$. These y-only dependent features are clearly visible in the data, *e.g.* Fig. 3 shows that higher viral loads are by far more rare than lower viral loads. In this case, the data-inferred $g(y)$ is non-zero and shows structure. Following the same reasoning, the most likely distribution for $h(x, y)$ in absence of data is identically-zero everywhere. Again, $h(x, y)$ will only be non-zero in case that the data triggers some coupling between x and y. Thus, the model favors independence of x and y (by favoring $h \equiv 0$) and the inferred density will be entangled in x and y only in the case the data exhibits this feature. We again assume a Matérn-kernel-shaped correlation structure for the Gaussian process g, with covariance parameters $p_g := (a_g, k_g, \gamma_g)$. We set the priors on p_g as for p_f and learn these parameters from the data as well. Since the typical length scales and amplitudes of the variations of $h(x, y)$ in x and y directions are not in principle a priori similar, we assume the covariance for h to be shaped by a direct product of individual Matérn kernels in the x and y directions. For their corresponding prior parameters, $p_h = (p_h^{(x)}, p_h^{(y)})$ with $p_h^{(i)} = (a_h^{(i)}, k_h^{(i)}, \gamma_h^{(i)})$ and $i \in \{x, y\}$, we use similar hyper-priors as before, i.e. as for p_f and p_g, respectively[3]. We call the ensemble of all these kernel parameters $p := (p_f, p_g, p_h)$. The details on how the Gaussian process for h with a Matérn kernel product covariance structure is set up is described in the Appendix, where we describe a multi-dimensional density estimator that is agnostic to causal directions.

[2] We choose the following priors on the signal parameters: $a_f = (0.3 \pm 0.1)$, $k_f = (5.0 \pm 3.0)\ \mathrm{yr}^{-1}$, and $\gamma_f = (-3.0 \pm 2.12)$, where the mean and standard deviation specify a Gaussian prior distribution for γ_f and log-normal distributions with the given mean and standard deviation for a_f and k_f.

[3] We choose the following priors on the remaining signal parameters: $a_g = (0.3 \pm 0.1)$, $k_g = (5.0 \pm 2.0)\ \mathrm{yr}^{-1}$, and $\gamma_g = (-2.75 \pm 2.12)$. For h we assume the prior correlation structures from p_f and p_g respectively for each axis.

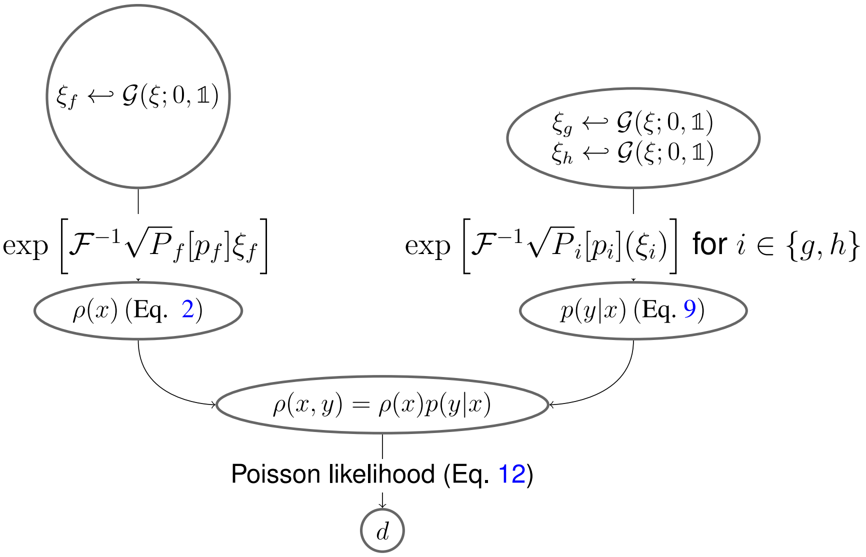
$\xi_f \hookleftarrow \mathcal{G}(\xi; 0, \mathbb{1})$
$\xi_g \hookleftarrow \mathcal{G}(\xi; 0, \mathbb{1})$
$\xi_h \hookleftarrow \mathcal{G}(\xi; 0, \mathbb{1})$
$\exp\left[\mathcal{F}^{-1}\sqrt{P}_f[p_f]\xi_f\right]$
$\exp\left[\mathcal{F}^{-1}\sqrt{P}_i[p_i](\xi_i)\right]$ for $i \in \{g, h\}$
$\rho(x)$ (Eq. 2)
$p(y|x)$ (Eq. 9)
$\rho(x,y) = \rho(x)p(y|x)$
Poisson likelihood (Eq. 12)
d

Figure 1: Graph structure of the causal model of age x and viral load y. $\mathcal{F}$ denotes the Fourier transform operator.

Finally, we normalize the conditional PDF

$$p(y|x) = \frac{e^{g(y)+h(x,y)}}{\int e^{h(\tilde{x},y)}\,\mathrm{d}\tilde{x}} \left(\int \frac{e^{g(\tilde{y})+h(x,\tilde{y})}}{\int e^{h(\tilde{x},\tilde{y})}\,\mathrm{d}\tilde{x}}\,\mathrm{d}\tilde{y}\right)^{-1} \tag{9}$$

such that the full model density reads

$$\varrho(x,y) = \varrho_0 e^{f(x)} \frac{e^{g(y)+h(x,y)}}{\int e^{h(\tilde{x},y)}\,\mathrm{d}\tilde{x}} \left(\int \frac{e^{g(\tilde{y})+h(x,\tilde{y})}}{\int e^{h(\tilde{x},\tilde{y})}\,\mathrm{d}\tilde{x}}\,\mathrm{d}\tilde{y}\right)^{-1}. \tag{10}$$

The assumed causal structure $x \rightarrow y$ is introduced in the model by the asymmetry between the roles of the x and y coordinates and the zero-centered Gaussian process priors on f, g, and h. Interchanging x and y leads to a model that follows the opposite causal direction $y \rightarrow x$. This allows to empirically distinguish these causal directions by calculating the model evidences for the two opposite scenarios, namely $x \rightarrow y$ and $y \rightarrow x$, as well as to test for $x \perp y$ by enforcing $h = 0$.

3.2 Likelihood

In order to construct the likelihood $\mathcal{P}(d|\varrho(\cdot,\cdot))$, we bin the data into a fine two dimensional grid over the x and y coordinates with 90×128 pixels, such that

$$n_{ij}(d) = \sum_{m=1}^{N} \int_{i\Delta x}^{(i+1)\Delta x} \mathrm{d}x \int_{j\Delta y}^{(j+1)\Delta y} \delta(x - x_m)\delta(y - y_m)\,\mathrm{d}y$$

contains the number of cases within the $(i,j)^{\text{th}}$ pixel of size $\Delta x = 1\,\text{yr}$ and $\Delta y \simeq 0.04\log_{10}(\text{viral RNA copies}_i)/\text{ml}$. These counts n_{ij} are then compared with the model's expectations

$$\begin{aligned} \lambda_{ij} \coloneqq \lambda_{ij}(\varrho) &= \int_{i\Delta x}^{(i+1)\Delta x} \mathrm{d}x \int_{j\Delta y}^{(j+1)\Delta y} \varrho(x,y)\,\mathrm{d}y \\ &\approx \Delta x \Delta y\; \varrho\bigg(\left(i + \frac{1}{2}\right)\Delta x, \left(j + \frac{1}{2}\right)\Delta y\bigg) \end{aligned} \tag{11}$$

via a Poisson likelihood

$$\mathcal{P}(d|\varrho) = \prod_{i,j} \frac{\lambda_{ij}^{n_{ij}}}{n_{ij}!} e^{-\lambda_{ij}}. \tag{12}$$

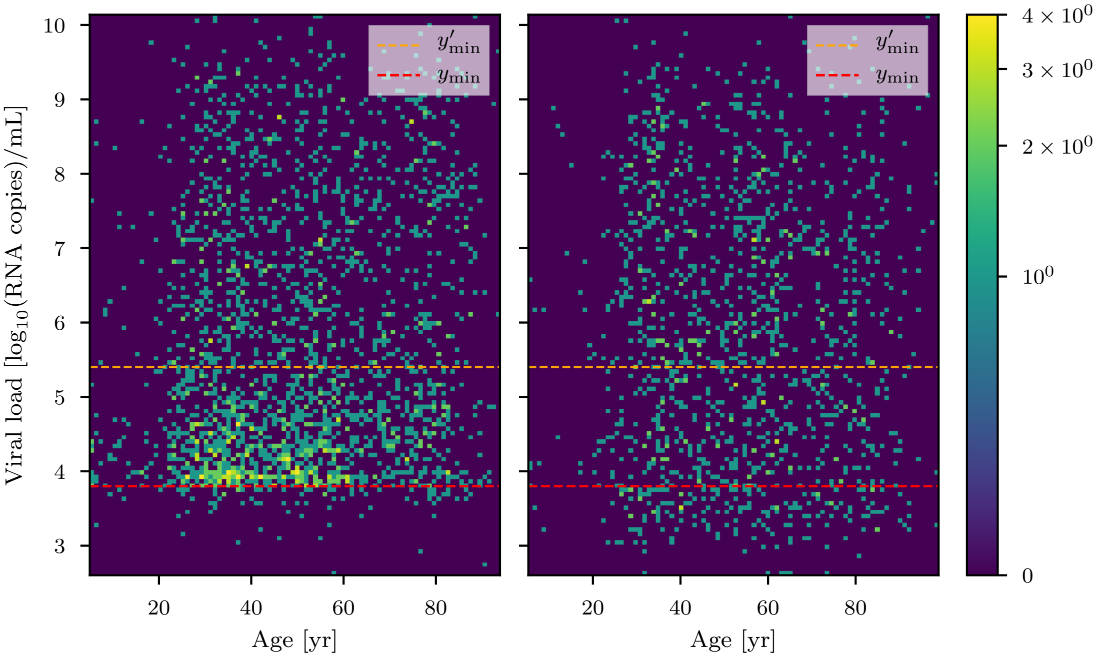

Figure 2: The cobas (left) and the LC 480 (right) datasets. The lower thresholds y_{min} and y'_{min} with which the data has been filtered are shown in red and orange, respectively.

3.3 Inference

The full model involving the data d as well as all the unknown quantities, which compose the signal vector $s := (f, g, h, p)$, reads

$$\begin{aligned} \mathcal{P}(d,s) &= \mathcal{P}(d|s)\,\mathcal{P}(s), \quad \text{where} \\ \mathcal{P}(d|s) &= \mathcal{P}(d|\varrho[f,g,h]) \quad \text{and} \\ \mathcal{P}(s) &= \mathcal{P}(f|p_f)\mathcal{P}(g|p_g)\mathcal{P}(h|p_h)\,\mathcal{P}(p_f)\,\mathcal{P}(p_g)\,\mathcal{P}(p_h). \end{aligned} \tag{13}$$

At this stage, we need to convert our causal model into an inference machine for the signal vector s. We do this conversion by reformulating the model in the language of information field theory [10, 11], transforming the coordinates of the signal vector $s = s(\xi)$ such that the prior on the new ξ coordinates becomes an uncorrelated Gaussian $\mathcal{P}(\xi) = \mathcal{G}(\xi, \mathbb{1})$ as described by Knollmüller and Enßlin [25]. We then implement the resulting model using the Python package `Numerical Information Field Theory` [NIFTy, 26–28] and finally use `NIFTy`'s implementation of Metric Gaussian Variational Inference [MGVI, 29] to approximate the posterior distribution in the new coordinates

$$\mathcal{P}(\xi|d) = \frac{\mathcal{P}(d|\xi)\mathcal{P}(\xi)}{\mathcal{P}(d)} \approx \mathcal{G}(\xi - \overline{\xi}_d, \Xi_d)$$

with a Gaussian which has posterior mean $\overline{\xi}_d$ and covariance Ξ_d, where the d suffix indicates the dataset used in the inference. This Gaussian posterior encodes the approximate result of the inference in the new coordinates. In order to translate this into the signal coordinates, we have to transform $\mathcal{P}(\xi|d)$ to $\mathcal{P}(s|d)$ using the relation $s = s(\xi)$. This relation is non-linear and the resulting PDF is neither Gaussian nor practical to obtain analytically. In order to evaluate moments from the posterior distributions of the desired quantities, MGVI provides ξ-samples drawn from the approximate Gaussian posterior $\xi \hookleftarrow \mathcal{G}(\xi - \overline{\xi}_d, \Xi_d)$. These ξ-samples can be converted via the coordinate transformation $s = s(\xi)$ into the signal space, where they represent (approximate) signal posterior samples. Making use of these posterior samples it is possible to calculate the posterior expectation values and model uncertainties of any desired quantity $q(s)$:

$$\begin{aligned} \overline{q} &:= \langle q(s) \rangle_{(s|d)} \approx \frac{1}{N_{\mathrm{s}}} \sum_{i=1}^{N_{\mathrm{s}}} q(s(\xi_i)) \\ \sigma_q^2 &= \left\langle (q(s) - \overline{q})^2 \right\rangle_{(s|d)} . \end{aligned} \tag{14}$$

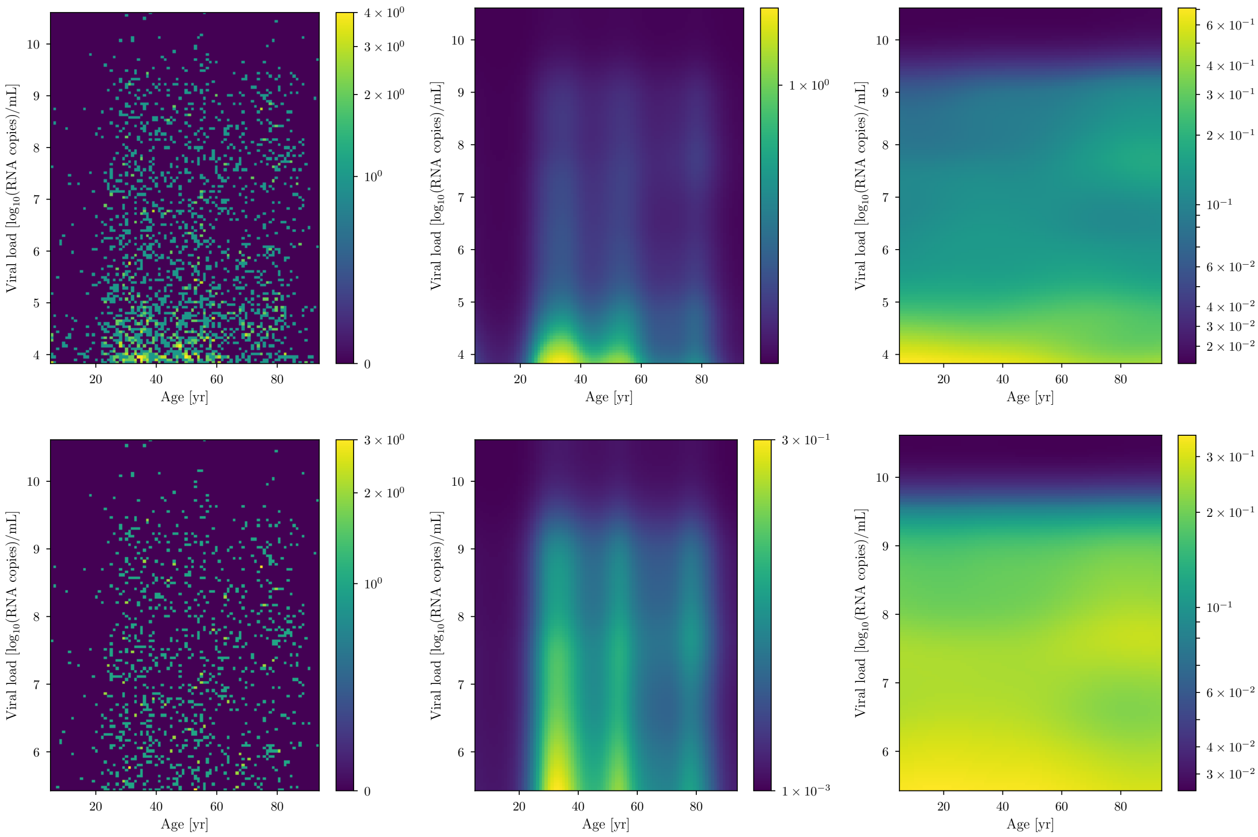

Figure 3: The cobas dataset (left), the reconstructed density distribution $\varrho(x, y)$ as a function of the age (x) and the viral load (y) in a logarithmic coloring scheme (middle) and the 2D conditional probability distribution of the viral load for two different data-filtering thresholds y_{min} (top) and y'_{min} (bottom).

Here, ξ_i denotes the i^{th} of the N_{s} drawn samples. In particular, the posterior mean of the conditional PDF $p(y|x)$ and of any quantity which can be calculated from $p(y|x)$ can thereby be obtained, as well as the resulting uncertainties characterized via their uncertainty dispersion. In general, these posterior signal samples will not follow Gaussian statistics because the transformation is typically non-linear. Furthermore, since MGVI is a variational inference approach, the calculated uncertainties will be slightly smaller compared to the ones given from the accurate posterior. However given the complexity and size of the model, we need to use an approximate inference method as MGVI. For details about this methodology, as well as extensive performance and accuracy tests, we refer to Knollmüller and Enßlin [29].

4 Data

We make use of RT-PCR viral load data collected from the Charité Institute of Virology and Labor in Berlin, Fig. 6 of [9]. The data was acquired with two different PCR instruments, Roche cobas 6800/8800 (cobas dataset, which we denote by d_{C} and is comprised of ≈ 2200 data points) and Roche LightCycler 480 II (LC 480 dataset, which we denote by d_{L} comprised of ≈ 1350 data points). In the following, we will show the difference between the two datasets. As can be seen from the count difference in the raw data plots as of Fig. 6 of [9] and in the histogramed data in Figs. 3 and 4, for low viral loads ($y \lesssim 5$ in units of $\log_{10}$(RNA copies/ml)), the LC 480 dataset shows a roughly uniform count distribution in the whole viral load domain. In contrast, the cobas dataset exhibits an increasing number of counts in the $y \in [2.0, 3.8]$ viral load domain followed by a descending trend in counts in the $y \in [3.8, 5.0]$ region. For this reason and in order to better understand the possible shortcomings of both instruments, we analyzed the data in two different ways.

Since the major differences between the datasets arise for viral load values $y \in [3.8, 5.0]$ we define two lower thresholds for the viral load ($y_{\text{min}} := 3.8$ and $y'_{\text{min}} := 5.4 \simeq \log_{10}(250000)$ in units of $\log_{10}$(viral RNA copies/ml)) and discard

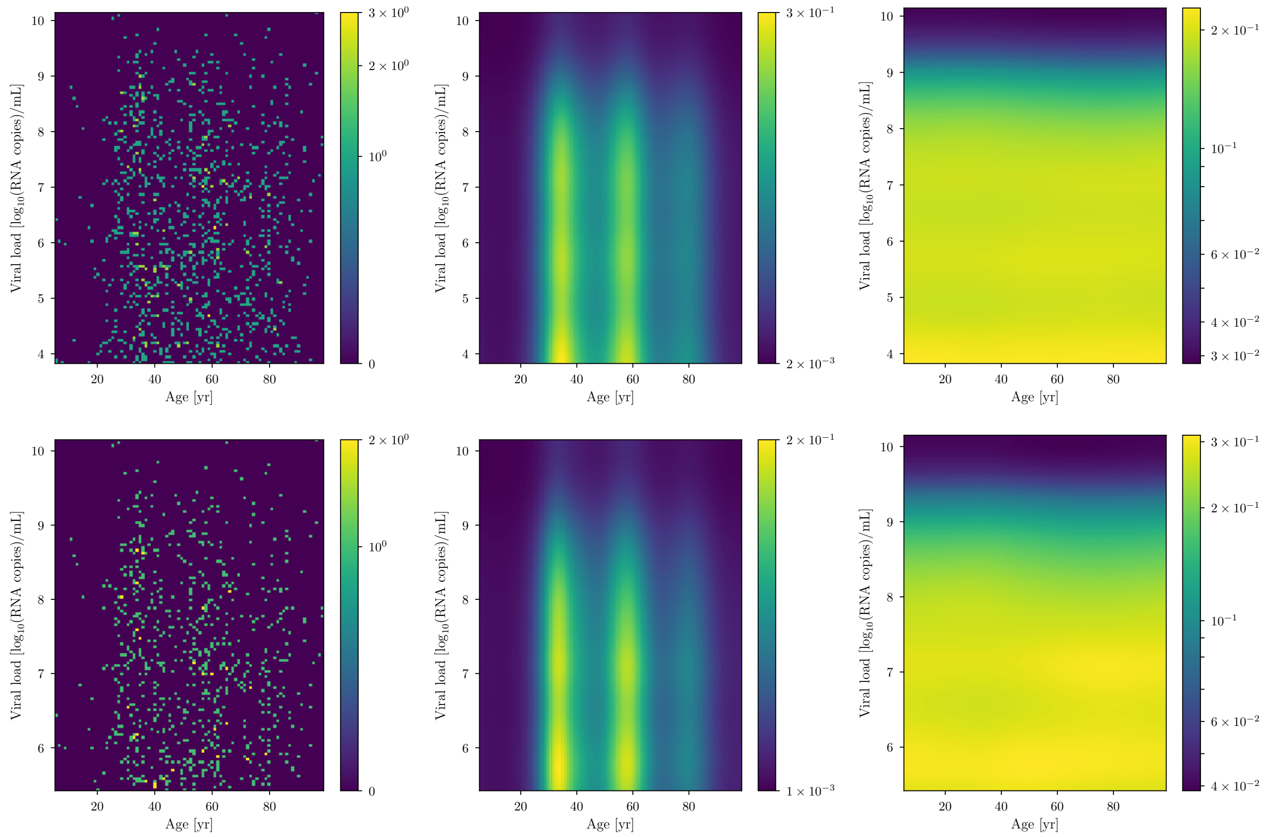

Figure 4: The same as in Fig. 3, but for the LC 480 dataset.

any data point for which the viral load is lower than y_{min} and y'_{min}, respectively. We set the lower threshold y_{min} at the value for which the number of counts of the cobas dataset is maximum (see Fig. 2). Below this threshold, the counts' density lowers dramatically. As for y'_{min}, the value of 250000 in units of viral RNA copies/ml indicates the threshold for the isolation of infectious virus in cell cultures at more than 5% probability as described by Wölfel et al. [30]. We then analyze the cobas and LC 480 datasets first neglecting the data below y_{min} threshold and then below y'_{min}. For the sake of simplicity, we will denote the cobas and LC 480 datasets with y_{min} and y'_{min} as a lower threshold as d_{C}, d'_{C}, d_{L}, and d'_{L}, respectively. This way we can highlight the differences between the two datasets and investigate possible sources of systematic errors in the viral load measurement instruments or different selection effects in the data collection process.

The datasets are not explicitly provided by the authors. Therefore, we acquired the data by means of a plot digitizer algorithm from Fig. 6 of [9]. Since the age coordinates are not labeled precisely, but only a rough interval $\Delta_{\text{age}} \sim 0 - 100\,\text{yr}$ is provided, we do not expect the acquired data points to be accurate – especially in the age domain. For this reason the obtained age axis could be affected by a global shift up to $5 - 10$ years in any direction. But, since our model is translation invariant, this kind of systematic bias in the data extraction process (an overall shift in the age values for all data points) does not affect the causal inference machinery. Nonetheless, the reader should keep this in mind when interpreting the results concerning the conditional PDF of the viral load given the age. Concerning the viral load coordinates, for which more precise units were given, the data should be regarded as more reliable. For our aim of building a method to analyze age and viral load data in a non-parametric and causal fashion providing uncertainty estimates, this level of accuracy is sufficient. The results we present from here on are given for the measured coordinate values without considering any uncertainty with respect to the real quantities (age, viral load). Nevertheless, we do not believe systematic or random error contributions in the data extraction procedure to significantly affect our results since the shapes of the learned distributions are translation invariant.

5 Results

In the following, we discuss the main results and their relevance for the infectivity of different age groups.

5.1 Age dependence of the viral load

For both lower thresholds and datasets d_{C}, d_{L}, we estimate the model parameters s by means of the MGVI algorithm implemented in `NIFTy`. The data and the correspondent reconstructed underlying density $\varrho(x, y)$ are shown in Figs. 3, 4.

A multi-modal age distribution is clearly visible as well as an overall decrease for growing viral loads in the all reconstructed densities. The conditional PDF $p(y|x)$ of the viral load y given the age x is shown in Fig. 5. In the conditional probability, the multi-modal age structure displayed by the density is not visible since it has been absorbed by $\varrho(x)$. What this effectively means is that age-only selection effects – e.g. testing one or many specific age groups more than others or demographics in general – have been modeled, and the resulting conditional probability distribution does not depend on such effects.

For all datasets and ages, a general descending trend in the viral load probability distribution is clearly visible. The reconstruction based on the d_{C} dataset exhibits significant differences in the viral load for different ages (Fig. 5, first panel). For infected patients approximately above the age of 60, the distribution exhibits a distinct maximum for viral loads of $y \simeq 8$ (in units of $\log_{10}$(RNA copies/ml)). Furthermore, we show that this feature is indeed triggered by the data, and is not just the result of an over-fit of sample noise. To do so, we apply a random permutation r to the viral load values in the d_{C} dataset, the only one that exhibits a possible $x \to y$ causal structure. We then analyze the randomized dataset $d^r_{\mathrm{C}} = \{(i, x_i, y_{r(i)})\}_{i=1}^{N}$ in the same way as seen for $d_{\mathrm{C}} = \{(i, x_i, y_i)\}_{i=1}^{N}$.The resulting conditional PDF $p^r(y|x)$ reconstructed from the randomized dataset (Fig. 5) does not exhibit any clear age-dependent structure in the viral load, indicating that the differences seen in the real data are not just a shot noise effect.

5.2 Causal directions and bias

We now focus on understanding the causal relations given by the interplay of the variables. The proposed causal model should allow to discriminate between all possible causal structures, namely $x \to y$, $y \to x$, and $x \perp y$. This will show that selection effects distorting the expected causal relation $x \to y$ are subdominant and that for the cobas dataset d_{C} we indeed see evidence for an age dependence of the viral load distribution, $p(y|x) \neq p(y)$.

First, we define the independent model, i.e. a model for which age and viral load are regarded as statistically independent variables. Such a model is built by setting a very tight zero-centered prior on h, hence removing it from the model for all practical purposes. We can then estimate the Bayesian evidence both for the causal – hence dependent – models ($x \to y$ and $y \to x$) and compare it with the evidence of the independent model ($x \perp y$). More precisely, we calculate the so called Evidence Lower Bound (ELBO) as a proxy for an evidence.[4] The $y \to x$ model is obtained by swapping the coordinates of the $x \to y$ model. The quantity of interest is then the logarithm of the evidence ratio of each causal model with the independent one,

$$\Delta E_{x \rightleftarrows y} = \log \frac{p(d|x \rightleftarrows y)}{p(d|x \perp y)}, \tag{15}$$

where $x \rightleftarrows y$ denotes either $x \to y$ or $y \to x$. $\Delta E_{x \to y}$ indicates the log evidence in favor of the causal model $x \to y$ with respect to the independent one and similarly $\Delta E_{y \to x}$ the one for $y \to x$. Comparing $\Delta E_{x \to y}$ with $\Delta E_{y \to x}$ also allows to discriminate between the two possible causal directions in the dataset.

The log-evidence ratio for the d_{C} dataset is $\Delta E_{\mathrm{C}, x \to y} = 4.6 \pm 1.0$ for d_{C}, which clearly favors the dependent model, but this value decreases to $\Delta E'_{\mathrm{C}, x \to y} = -1.5 \pm 1.0$ when considering d'_{C}, hence y'_{min} as a lower threshold. We highlight that a log-evidence difference between the two compared models of 1 unit corresponds to a factor of $e \approx 2.7$ for the Bayesian odds ratio between the two. Thus, $\Delta E_{\mathrm{C}, x \to y} = 4.6 \pm 1.0$ implies given equal model priors, $p(x \to y) = p(x \perp y)$, a posterior model odds ratio of $p(x \to y|d) : p(x \perp y|d) = e^{3.5 \pm 0.7} \approx 99.5^{[270.4]}_{[36.6]}$ in favor of a causal dependence between viral load and age for the cobas dataset with the lower threshold y_{min}.

For the opposite causal direction we get $\Delta E_{\mathrm{C}, y \to x} = -0.2 \pm 1.1$, which shows that there is no strong $y \to x$ structure in the data. For the LC dataset d_{L} these evidence differences with respect to the independent model become $\Delta E_{\mathrm{L}, x \to y} = -4.6 \pm 1.0$ and $\Delta E'_{\mathrm{L}, x \to y} = -3.4 \pm 1.0$ respectively for the two thresholds. Therefore the independent

[4] Using the posterior uncertainty covariance Ξ as well as the posterior samples provided by MGVI, we can compute the ELBO [31] for each model. This is lower than the exact logarithm of the evidence by the information difference (as measured in nits) between the exact and approximated posterior of a model. If not too much information is lost in MGVI, the ELBO should be a good approximation of the exact log evidence. Furthermore, we can assume that deviations from the exact log evidence should be similar among different models, thereby reducing the effect of the MGVI approximations on differences between log-evidences. Thus, we can use the ELBO as a good proxy for the log model evidence ratios of similar models. The stochastic sampling steps performed in order to estimate the ELBO introduce a sampling uncertainty. This uncertainty can be in principle reduced by taking more samples, at the expense of larger computational costs. We state this numerical one-sigma uncertainty for all MGVI and ELBO based log-evidences.

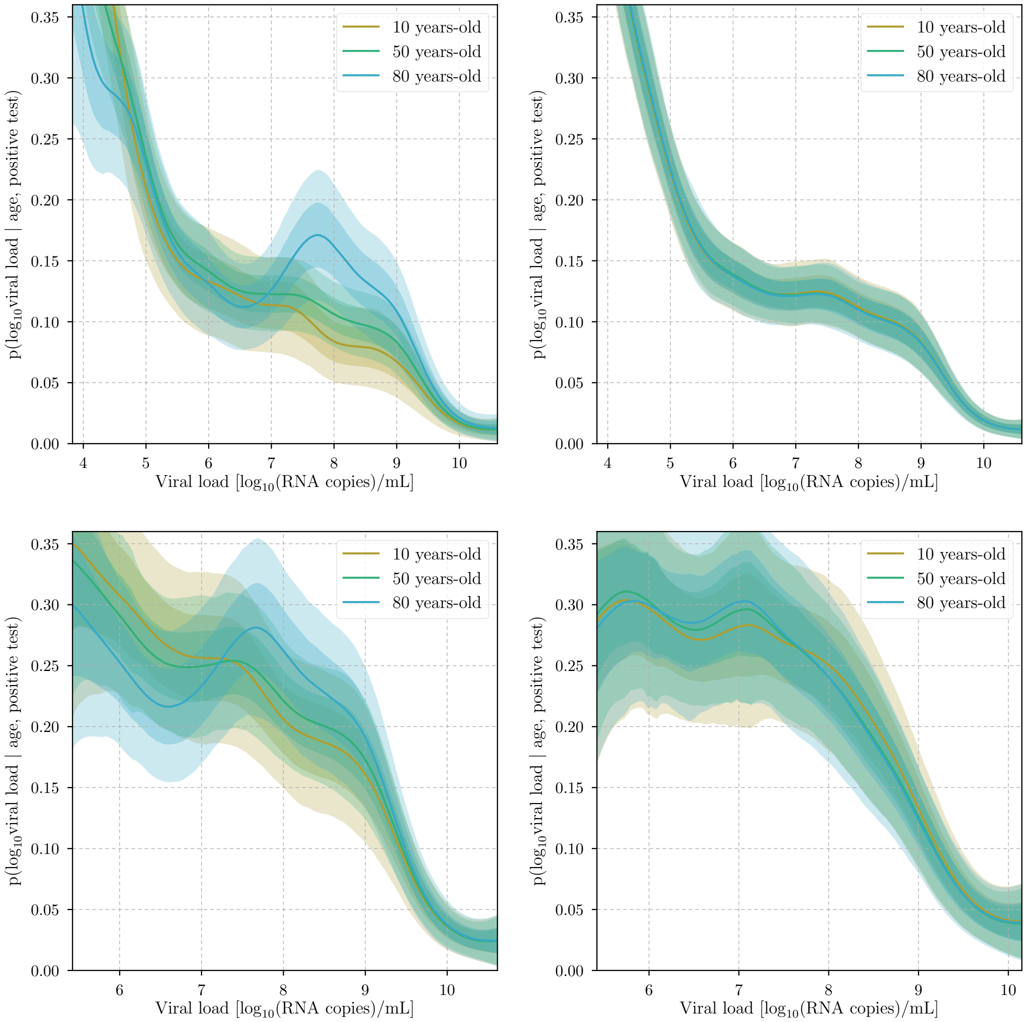

Figure 5: Viral load conditional PDF $p(y|x)$ for specific ages $x \in \{10, 50, 80\}$. Panel one and two (top) display the results for the cobas dataset d_C and for the randomized cobas dataset d_C respectively. The latter serves as a null test, since the randomization erases any (causal) relation between the age (x) and the viral load (y)other than shot noise. Panel three and four (bottom) show $p(y|x)$ for the cobas dataset d'_C and for the LC 480 dataset d'_L respectively, both with the higher viral load threshold y'_min. The shaded regions represent 1σ and 2σ uncertainty contours of the approximate posterior.

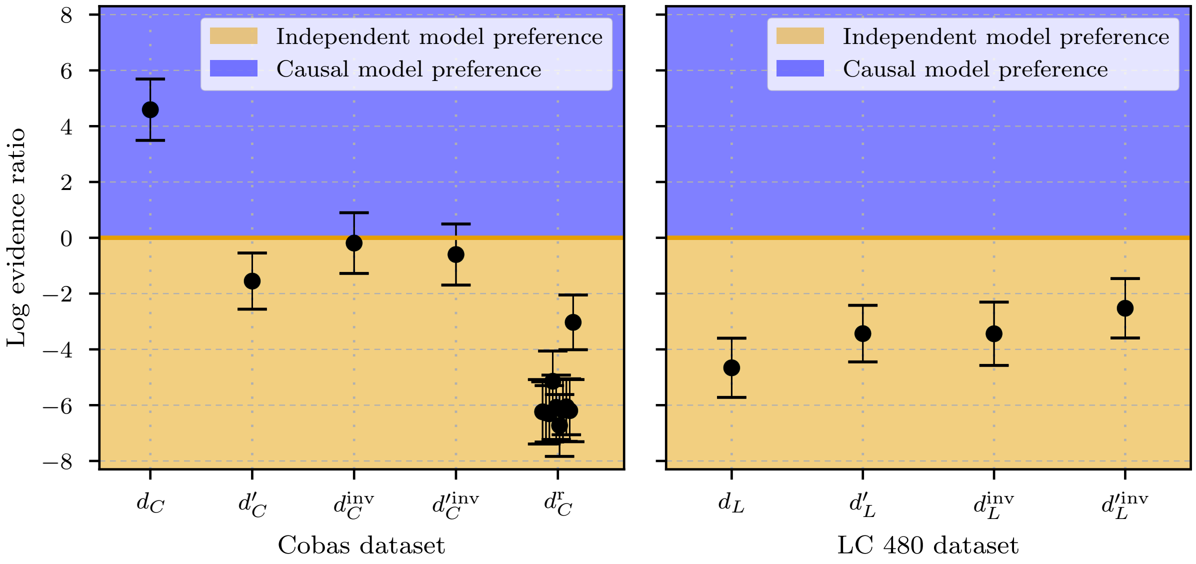

Figure 6: Natural logarithm of the evidence ratio with respect to the correspondent independent model for the $x \to y$ cobas datasets $d_{\rm C}$ and $d'_{\rm C}$, for the $y \to x$ causal model $d_{\rm C}^{\rm inv}$ and $d'\,^{\rm inv}_{\rm C}$(left panel). The log-evidence ratios labeled with d_C^r display the (y-)randomized datasets (for 10 different realizations). The same is also shown for the LC 480 dataset (right). The error bars represent the numerical uncertainty associated to the stochastic estimate of the ELBO. A positive logarithm of the evidence ratio denotes a preference for the given causal model with respect to the correspondent independent model and vice versa.

model is favored in both cases. This shows that the cobas ($d'_{\rm C}$) and LC 480 ($d'_{\rm L}$) datasets are in agreement for viral loads which are higher than $y'_{\rm min}$, but in case we include the data lying in the viral load region $y \in [y_{\rm min}, y'_{\rm min}]$, the causal age-dependent structure becomes visible in $d_{\rm C}$ and is not negligible anymore.

It is known that model evidences can vary strongly with different data realizations. Moreover, in order to calculate the evidences, we invoked approximations and stochastic calculation steps. Thus, proper null-tests are required in order to validate and calibrate the evidence ratio calculation. We provide such null-tests by repeating the data randomization step described above several times, thereby producing many randomized datasets. By construction, these dataset should not exhibit any causal structure. Indeed, for the 10 randomized-dataset realizations performed, we find much lower log-evidence ratios between dependent (causal) and independent models with respect to the ones found for the original cobas dataset $d_{\rm C}$. The average difference between these tests is $\langle \Delta E_{{\rm random},x\to y}\rangle_{\rm randomizations} = -6.0 \pm 1.0$. Since none of the randomized dataset realizations reaches comparably high log-evidence ratios with respect to the independent model, all these findings support robustly the argument that the dependent model is a more suited description of the $d_{\rm C}$ dataset.

The results of the evidence calculations are displayed in Fig. 6. This figure also indicates that the independent model interpretation of the data is favored for all other datasets and threshold combinations (except for $d_{\rm C}$), since the evidence for the independent model is always higher.

These contradicting results for the two datasets (or thresholds) $d_{\rm C}$ and $d'_{\rm C}$ might have several possible explanations. First, they could be caused by a potential accuracy loss of the PCR devices below certain viral load values, as suggested for the cobas dataset $d_{\rm C}$ below $y'_{\rm min}$ in [9], hence for the Roche cobas 6600/8800 PCR system. It could also mean that the opposite is happening and the Roche LC 480 PCR device is less sensitive than the Roche cobas 6600/8800 in the $y \in [3.8, 5.0]$ region. This possibility would be supported by the fact that the cobas dataset exhibits a clear age dependent structure in such viral load region, but the (swab) data processed by the cobas PCR device contains no information on the patients' ages. Hence it would be surprising that a systematic effect in the measurements could introduce an age dependence on the viral load distribution. Furthermore, this pattern – that older patients exhibit higher viral loads – is plausible from a medical perspective.

Nevertheless, we cannot exclude the possibility that selection effects have been introduced in the data. We have already shown that "viral load causing age" effects ($y \to x$) are subdominant. Nevertheless, selection effects could still have

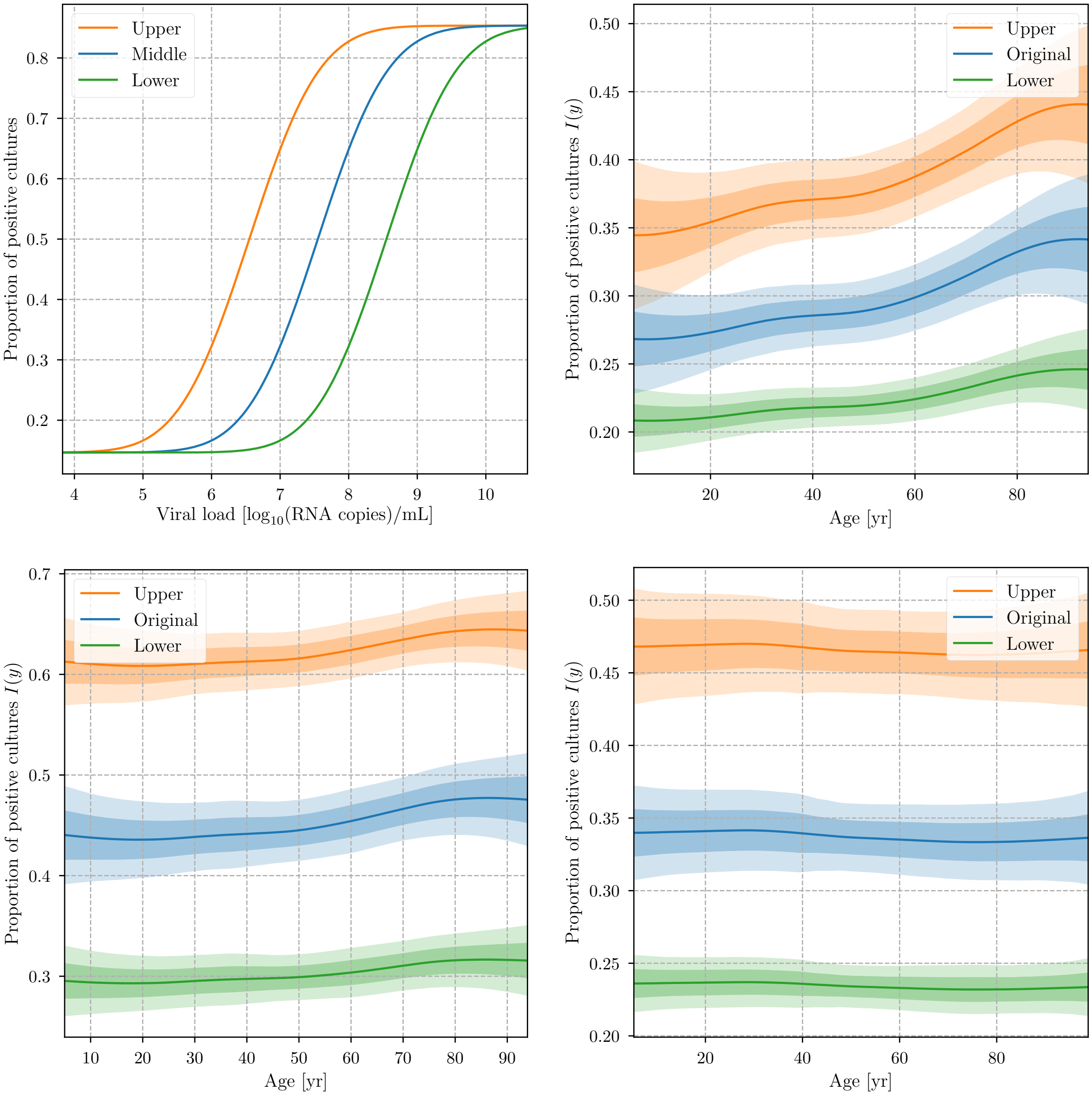

Figure 7: $I(y)$ obtained from Fig. 1g in [30] and fit with a probit function (top left, Original), then projected for different patients' ages for the cobas dataset d_{C} (top right), d'_{C} (bottom left) and for the LC 480 dataset d_{L} (bottom right). The upper and lower curves are obtained by translating $I(y)$ of one unit in viral load as in $I(y-1)$ and $I(y+1)$ in order to give upper and lower bounds similar as those shown in Fig. 1g in [30]. The shaded regions show 1σ and 2σ uncertainty contours of the approximate posterior.

been introduced for instance by collecting age and viral load subsamples from a "viral-load biased" population sample. This could happen e.g. if symptomatic patients would have been predominantly tested. Since children are less likely to show symptoms than adults, the sample would then include mainly those children with higher viral loads.

And finally, a combination of such competing effects could have affected the results and thereby imprinted a spurious age dependence to the viral-load distribution. Given only the statistical data in our possession, this possibility cannot be ruled out completely.

5.3 Impact on infectivity

After having established a potential age difference in the viral load distribution for d_{C}, we investigate whether this difference - if real - would be relevant for the infection dynamics. For this purpose, we need to link the viral load to the infectivity $I(y)$ of the virus, i.e. the probability of transmitting the infection. Infectivity can be measured in different ways. In our analysis, we choose the projected virus isolation success based on probit distribution described in [30] as a proxy for infectivity. This represents the infection success rate for cell cultures exposed to saliva with different viral load y and can be seen as the blue curve in Fig. 7, labeled with "Original". We will from now on refer to this parameter as "infectivity proxy" and indicate it with $I(y)$.

The projected infectivity as a function of the age is then given by the expectation value of the infectivity parameter over the conditional PDF $p(y|x)$

$$I(x) \coloneqq \langle I(y) \rangle_{p(y|x)} = \int I(\tilde{y}) p(\tilde{y}|x) \, \mathrm{d}\tilde{y} \, . \tag{16}$$

The result, together with the uncertainties resulting from our PDF modeling are shown in Fig. 7. No relative differences larger than 0.18 for the (projected) infectivity of the different age groups is found, with typical values of $I(x) \approx 0.3$ for all datasets. This means that at most a 50% difference in infectivity due to different viral load between different age groups - but more likely a smaller one - should be expected.

In order to characterize the uncertainty resulting from our $I(y)$ modeling, due to the uncertainty of the original determination of this function and due to the uncertainty in the identification of viral loads with different instruments, we repeat the analysis while shifting the original $I(y)$ curve by one order of magnitude upwards and downwards in y. The resulting maximal relative difference in the infectivity of the different age groups is $\simeq 0.3$. Thus, although our model allows us to show that infectivity exhibits an age dependence if the $d_{\mathbf{C}}$ dataset with y_{min} provides a valid picture, the viral load differences between different age groups though are not strong enough to impact on the infection dynamics at a level that justifies regarding any age group as noninfectious or even significantly less infectious.

6 Conclusions

In order to investigate the controversial results reported in the literature, we developed a causal model to assess the dependence of viral loads of patients infected with COVID-19 on age. The model is very flexible and generic and can therefore be used in future epidemiological studies as well as in completely different contexts. We provide the source code under an open source license[5] for usage in further studies and applications. As a side product, we also developed a causal-direction-agnostic density estimator, which is described in more detail in the Appendix.

Using our novel method to model causal relations non-parametrically, we have reanalyzed the SARS-CoV-2 age and viral load data presented in [9].In doing so, we have found statistically significant differences in the viral load distribution of different age groups when regarding the cobas dataset d_{C} for viral loads within the interval of $10^{3.8}$ to $10^{5.0}$ in units of viral RNA copies/ml of sample or entire swab specimen as reliable. These differences become irrelevant if this region is ignored in the analysis.

We cannot completely exclude that selection effects in the data-collection process may have introduced an apparent causal relation between viral load and age, but the observed trend – a statistically-significant increase in the viral load with age – fits with the generally accepted notion that the immune system response gets weaker with age. Assuming this trend to be real, we showed, however, that its expected impact on the infectivity of different age groups is at most moderate. For this reason we cannot exclude any age group from being considered as a potentially significant source of infection.

The region of the cobas dataset relevant for this trend is described in [9] as containing an artifact, suggesting that the correct interpretation of the data is that viral load, hence infectivity, is predominantly age independent. Here, we want to point out that other studies on the age dependence of the viral load present in the literature [15, 18] make the opposite claim. Moreover, in their most recent publication, Jones et al. [24] acknowledge an age dependence of the viral load. This dependence is quantitatively similar to the one we have detected with our method. Furthermore, the causal evidence tests presented in Sect. 5 favor considering the age dependence of the viral load as a real effect and not just as an artifact. These tests also disfavor the reverse-causal-direction model (here: that the viral load of a patient "causes" its age), which would indicate that strong selection effects have affected the data-collection process. We introduced these tests as a new tool to detect potential systematic effects in similar datasets.

[5] https://gitlab.mpcdf.mpg.de/ift/public/causal_age_viral_load_model.

Acknowledgments

We thank Matthia Sabatelli (Montefiore Institute, University of Liège) for feedback on preliminary versions of this work and the Information Field Theory group at the Max Planck Institute for Astrophysics for the fruitful discussions.

References

[1] Philippe Colson et al. "Children account for a small proportion of diagnoses of SARS-CoV-2 infection and do not exhibit greater viral loads than adults". In: *European Journal of Clinical Microbiology & Infectious Diseases* 39.10 (2020), pp. 1983–1987.

[2] Shufa Zheng et al. "Viral load dynamics and disease severity in patients infected with SARS-CoV-2 in Zhejiang province, China, January-March 2020: retrospective cohort study". In: *Bmj* 369 (2020).

[3] Elisabet Pujadas et al. "SARS-CoV-2 viral load predicts COVID-19 mortality". In: *The Lancet Respiratory Medicine* 8.9 (2020), pp. 831–934. DOI: 10.1016/S2213-2600(20)30354-4.

[4] Jesse Fajnzylber et al. "SARS-CoV-2 viral load is associated with increased disease severity and mortality". In: *Nature Communications* 11.1 (2020), p. 5493. DOI: 10.1038/s41467-020-19057-5.

[5] Helena C Maltezou et al. "Association Between Upper Respiratory Tract Viral Load, Comorbidities, Disease Severity, and Outcome of Patients With SARS-CoV-2 Infection". In: *The Journal of Infectious Diseases* (Jan. 2021). jiaa804. DOI: 10.1093/infdis/jiaa804. eprint: https://academic.oup.com/jid/advance-article-pdf/doi/10.1093/infdis/jiaa804/36157672/jiaa804.pdf. URL: https://doi.org/10.1093/infdis/jiaa804.

[6] Lael M. Yonker et al. "Pediatric severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): clinical presentation, infectivity, and immune responses". In: *The Journal of pediatrics* 227 (2020), pp. 45–52.

[7] Tom Jefferson et al. "Viral cultures for COVID-19 infectivity assessment. Systematic review". In: *medRxiv* (2020).

[8] H. Kawasuji et al. "Transmissibility of COVID-19 depends on the viral load around onset in adult and symptomatic patients". In: *PLoS ONE* 15.12 (2020). DOI: 10.1371/journal.pone.0243597.

[9] Terry C Jones et al. "An analysis of SARS-CoV-2 viral load by patient age". In: *MedRxiv* (2020).

[10] Torsten A. Enßlin, Mona Frommert, and Francisco S. Kitaura. "Information field theory for cosmological perturbation reconstruction and nonlinear signal analysis". In: *Physical Review D* 80.10 (2009), p. 105005.

[11] Torsten A. Enßlin. "Information theory for fields". In: *Annalen der Physik* (2018), p. 1800127.

[12] R. H. Leike, M. Glatzle, and T. A. Enßlin. "Resolving nearby dust clouds". In: *Astronomy & Astrophysics* 639 (2020), A138.

[13] Philipp Arras et al. "The variable shadow of M87". In: *arXiv preprint arXiv:2002.05218* (2020).

[14] Maximilian Kurthen and Torsten Enßlin. "A Bayesian Model for Bivariate Causal Inference". In: *Entropy* 22.1 (2020), p. 46.

[15] Waleed H. Mahallawi et al. "Association of Viral Load in SARS-CoV-2 Patients With Age and Gender". In: *Frontiers in Medicine* 8 (2021), p. 39. DOI: 10.3389/fmed.2021.608215. URL: https://www.frontiersin.org/article/10.3389/fmed.2021.608215.

[16] Damien Jacot et al. "Viral load of SARS-CoV-2 across patients and compared to other respiratory viruses". In: *Microbes and infection* 22.10 (2020), pp. 617–621.

[17] Steven Kleiboeker et al. "SARS-CoV-2 viral load assessment in respiratory samples". In: *Journal of Clinical Virology* 129 (2020), p. 104439.

[18] Sjoerd Euser et al. "SARS-CoV-2 viral load distribution in different patient populations and age groups reveals that viral loads increase with age". In: *medRxiv* (2021). DOI: 10.1101/2021.01.15.21249691. eprint: https://www.medrxiv.org/content/early/2021/01/17/2021.01.15.21249691.full.pdf.

[19] Rosa Costa et al. "Upper respiratory tract SARS-CoV-2 RNA loads in symptomatic and asymptomatic children and adults". In: *medRxiv* (2021).

[20] Taylor Heald-Sargent et al. "Age-related differences in nasopharyngeal severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) levels in patients with mild to moderate coronavirus disease 2019 (COVID-19)". In: *JAMA pediatrics* 174.9 (2020), pp. 902–903.

[21] Helena C Maltezou et al. "Children and adolescents with SARS-CoV-2 infection: epidemiology, clinical course and viral loads". In: *The Pediatric Infectious Disease Journal* 39.12 (2020), e388–e392.

[22] Arnaud G L'Huillier et al. "Culture-competent SARS-CoV-2 in nasopharynx of symptomatic neonates, children, and adolescents". In: *Emerging infectious diseases* 26.10 (2020), p. 2494.

[23] Philip Zachariah et al. “Symptomatic infants have higher nasopharyngeal SARS-CoV-2 viral loads but less severe disease than older children”. In: *Clinical Infectious Diseases* 71.16 (2020), pp. 2305–2306.

[24] Terry C. Jones et al. “Estimating infectiousness throughout SARS-CoV-2 infection course”. In: *Science* (2021). ISSN: 0036-8075. DOI: 10.1126/science.abi5273. eprint: https://science.sciencemag.org/content/early/2021/05/24/science.abi5273.full.pdf. URL: https://science.sciencemag.org/content/early/2021/05/24/science.abi5273.

[25] Jakob Knollmüller and Torsten A. Enßlin. “Encoding prior knowledge in the structure of the likelihood”. In: *arXiv e-prints*, arXiv:1812.04403 (Dec. 2018), arXiv:1812.04403. arXiv: 1812.04403 [stat.ML].

[26] M. Selig et al. “NIFTy - Numerical Information Field Theory. A versatile Python library for signal inference”. In: *Astron. & Astrophys.* 554, A26 (June 2013), A26. DOI: 10.1051/0004-6361/201321236. arXiv: 1301.4499 [astro-ph.IM].

[27] Theo Steininger et al. “NIFTy 3 - Numerical Information Field Theory: A Python Framework for Multicomponent Signal Inference on HPC Clusters”. In: *Annalen der Physik* 531.3 (Mar. 2019), p. 1800290. DOI: 10.1002/andp.201800290.

[28] Philipp Arras et al. *NIFTy5: Numerical Information Field Theory v5*. Mar. 2019. ascl: 1903.008.

[29] Jakob Knollmüller and Torsten A. Enßlin. “Metric Gaussian Variational Inference”. In: *arXiv e-prints*, arXiv:1901.11033 (Jan. 2019), arXiv:1901.11033. arXiv: 1901.11033 [stat.ML].

[30] Roman Wölfel et al. “Virological assessment of hospitalized patients with COVID-2019”. In: *Nature* 581.7809 (2020), pp. 465–469. DOI: 10.1038/s41586-020-2196-x.

[31] Andrija Kostić et al. “Bayesian Causal Inference with Information Field Theory”. In: *Manuscript in preparation.* (May 2021).

[32] Wei-Chia Chen, Ammar Tareen, and Justin B. Kinney. “Density Estimation on Small Data Sets”. In: *Phys. Rev. Lett.* 121.16, 160605 (Oct. 2018), p. 160605. DOI: 10.1103/PhysRevLett.121.160605. arXiv: 1804.01932 [physics.data-an].

[33] Bernard W Silverman. *Density estimation for statistics and data analysis*. Vol. 26. CRC press, 1986.

[34] Simon J. Sheather. “Density Estimation”. In: *Statistical Science* 19.4 (2004), pp. 588–597. ISSN: 08834237. URL: http://www.jstor.org/stable/4144429.

[35] Qiao Liu et al. “Density estimation using deep generative neural networks”. In: *Proceedings of the National Academy of Sciences* 118.15 (2021).

[36] Thomas S. Ferguson. “A Bayesian Analysis of Some Nonparametric Problems”. In: *The Annals of Statistics* 1.2 (1973), pp. 209–230. DOI: 10.1214/aos/1176342360. URL: https://doi.org/10.1214/aos/1176342360.

[37] Peter Müller. *Bayesian nonparametric data analysis*. Cham, Switzerland: Springer, 2015. ISBN: 978-3-319-18968-0.

[38] Andrew Gelman. *Bayesian data analysis*. Boca Raton: CRC Press, 2014. ISBN: 978-1439840955.

[39] Justin B. Kinney. “Estimation of probability densities using scale-free field theories”. In: *Phys. Rev. E* 90 (1 July 2014), p. 011301. DOI: 10.1103/PhysRevE.90.011301. URL: https://link.aps.org/doi/10.1103/PhysRevE.90.011301.

[40] Justin B. Kinney. “Unification of field theory and maximum entropy methods for learning probability densities”. In: *Phys. Rev. E* 92 (3 Sept. 2015), p. 032107. DOI: 10.1103/PhysRevE.92.032107. URL: https://link.aps.org/doi/10.1103/PhysRevE.92.032107.

[41] Marc G. Genton. “Classes of Kernels for Machine Learning: A Statistics Perspective”. In: *J. Mach. Learn. Res.* 2 (Mar. 2002), pp. 299–312. ISSN: 1532-4435.

[42] Philipp Arras et al. *M87* in space, time, and frequency*. 2021. arXiv: 2002.05218 [astro-ph.IM].

A Matérn-kernel density reconstruction

We aim to infer an unknown distribution from a random realization of discrete data. In the following, we present a general-purpose density estimator that serves this scope. Using a fully Bayesian framework, we can propagate uncertainties through each inference step and extract information from the correlation structure inherent to the data. We provide our method as open-source software.

According to Chen, Tareen, and Kinney [32], the problem of extracting a smooth density function from a limited set of data samples is a challenging and well-known problem in statistical learning and data analysis. The most common ad-hoc methods to empirically derive (probability) densities from data usually involve histograms or Kernel Density Estimation (KDE) [33, 34]. These methods do not infer the smoothness of the learned (probability) density’s

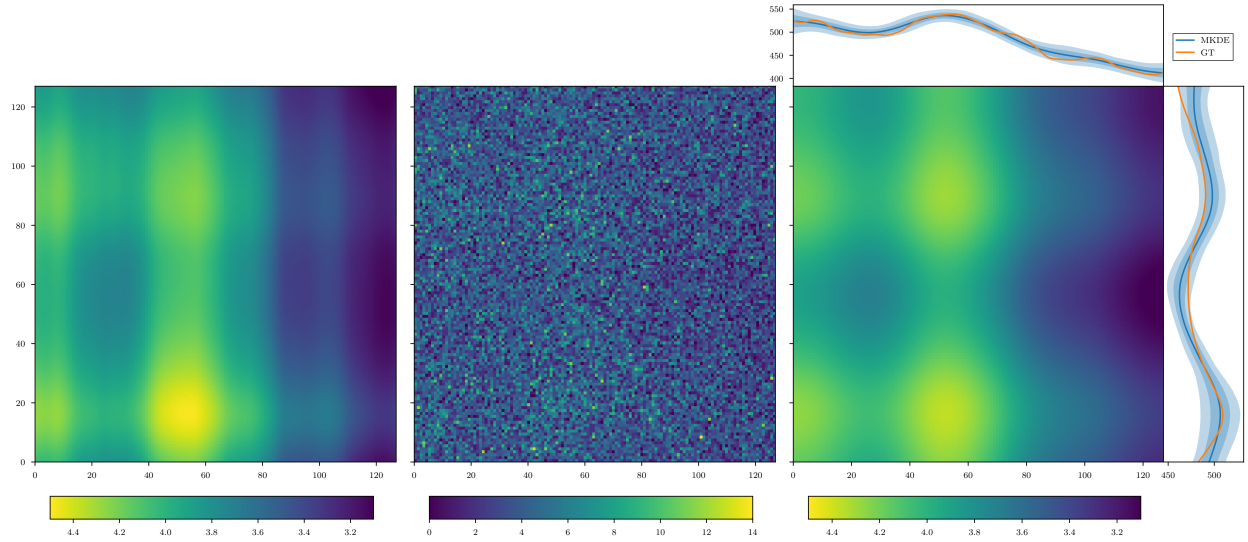

Figure 8: Example of MKDE performance. Left: random realization of a smooth probability density with independent covariance structure in the x and y direction (ground truth, GT). Middle: Poissonian counts drawn from the GT. Right: MKDE reconstruction from the counts. In the same panel, we also show the marginals of the reconstructed density (blue lines), together with their one and two-sigma uncertainty estimates (shaded areas in blue) against the marginals of the GT (orange lines).

correlation structure and are thus prone to reconstructing unphysical densities. Other methods make use of neural networks (see for example Liu et al. [35]) or restrict the density to specific functional forms (see e.g. Dirichlet Process Mixture Model, or DPMM [36–38]). Another approach is to use smooth priors and infer the level of smoothness of the reconstructed density from data via maximum entropy (Density Estimation using Field Theory, or DEFT [39, 40]). Chen, Tareen, and Kinney [32] propose an interesting information-theoretical-based modification of DEFT, although it only effectively works in one dimension. Finally, another very commonly used and effective solution to the density estimation problem is the one given by deep neural networks. In a recent work, Liu et al. [35] propose a generative adversarial networks (GAN) which is particularly effective in high dimensions. We refer to their work for comparison with similar neural-network-based approaches.

Most of these approaches lack a robust estimate for uncertainties or specify none at all. In the hope of addressing these shortcomings, we propose our novel general-purpose density reconstruction method, which we will refer to as Matérn Kernel Density Estimator (MKDE). This method is very general, works for a generic n-dimensional space, and therefore applies to many different contexts and fields. We presented one paradigmatic example of its many possible applications in Sec. 3.1. In this example, we used MKDE to reconstruct the continuous distributions of the ages and viral loads of patients infected with Covid-19 from age-and-viral-load data samples collected within the general population.

MKDE is capable of reconstructing a smooth density distribution underlying an – even limited – discrete dataset. We achieve this result under the hypotheses that the data points are drawn from the underlying density through a Poisson process and that the reconstructed density is a sufficiently smooth function. Since we expect the inferred (probability) density to be strictly positive and to vary on logarithmic scales, we choose the log-normal model

$$\varrho(x) = e^{s(x)}, \tag{17}$$

where $s(x)$ is the natural logarithm of the density. In the case of multi-dimensional data, both the density $\varrho(x)$ and its logarithm $s(x)$ in Eq. 17 are functions defined over n-dimensional vector spaces, with n being the dimension of the data space (e.g. space, time, age, viral load, ...). For the sake of simplicity, we will initially show how a one-dimensional density is reconstructed, emphasizing how to generalize to the multi-dimensional case only when this generalization is non-trivial.

Smoothness is a common and ubiquitous assumption when dealing with physical data. To fulfill this assumption in the MKDE, we parametrize the two-point correlation structure of the Gaussian process that determines the value of the log-density $s(x)$ at each point and for each dimension with a Matérn kernel. This kernel is a very flexible choice [41]. Moreover, it is very well suited to represent a priori homogeneous covariance structures like the ones we want to model. We define the Gaussian process $s := s(x) = (s_x)_{x\in\mathbb{R}} \hookleftarrow \mathcal{P}(s) = \mathcal{G}(s, S)$ on the one-dimensional position space $\mathbb{R}$.

This process has a homogeneous covariance structure $S_{xx'} = C_s(x - x')$ that can be efficiently represented in Fourier space thanks to the Wiener-Khinchin theorem. Invoking this theorem, we can use the power spectrum $P_s(k)$ to fully determine the Fourier transform of the two-point correlation function $C_s(x - x')$ for any stationary and statistically homogeneous process, and in particular for the Gaussian process s.

In order to draw prior samples from the Gaussian field s, we choose a standardized coordinate system $\hat{\xi}$. We then transform the standard normally distributed parameters $\xi = (\xi_k)_k$, where $k \in \mathbb{N}$ is the Fourier-space index according to the mapping

$$s = \mathcal{F}^{-1} A\, \xi \quad \text{with } A := \mathrm{diag}(\sqrt{P_s}). \tag{18}$$

Here, $\mathcal{F}$ represents the Fourier transform operator and A the amplitude operator in Fourier space. The amplitude operator encodes the Matérn-kernel correlation structure, parametrized with

$$A_{kk'} = 2\pi\, \delta(k - k') \frac{a_s}{\left[1 + k^2/k_0^2\right]^{-\gamma_s/4}}, \tag{19}$$

where a_s is a scale factor which accounts for the standard deviation in position space, k_0 is the magnitude of the characteristic correlation-length wavevector, and γ_s is the spectral index of the power spectrum. We assume a_s and k_0 to be a priori log-normally distributed since we expect strictly positive variations of the possible power spectra on a logarithmic scale. Similarly, we choose γ_s to be normally distributed, since the spectral index could in principle also be negative. We additionally introduce volume factors to ensure that the model parameters are intensive with respect to volume, i.e. they do not depend on the volume in position space.[6]

For higher-dimensional data, we expect the correlation structure along each axis (or dimension) to be a priori independent from the others. These different axes could in fact have very different meanings (and units), as they could represent – for instance – space and time, temperature, pressure, and volume, age and viral load (as seen in Sec. 3.1), or a different combination of these and other continuous quantities. Therefore, in the n-dimensional data space we can decompose the amplitudes of the correlation structure

$$A_{k,q,\ldots,n} = \bigotimes_{i \in \{k,q,\ldots,n\}} A_i \tag{20}$$

along each independent axis, each modeled by an individual amplitude operator A_i, for $i \in \{k, q, \ldots, n\}$. The zero modes of the individual axes must be treated separately in order to avoid degeneracy. Thus, in the proposed model the zero mode is shared among all directions and inferred independently through an a priori strictly-positive and uniformly-distributed parameter α.[7] For more details on the zero mode degeneracy and factorizing power spectra, we refer to Arras et al. [42].

At this stage, we can summarize all the parameters that we have introduced for each independent axis with the scalar-valued parameters α, a_s, k_0, and γ_s and the vector-valued ξ_k. We set broad priors on these parameters[8] and learn them using Metric Gaussian Variational Inference (MGVI). For details on the inference of posterior estimates for the MKDE parameters through MGVI and their uncertainty quantification, we refer to Sec. 3.3. Fig. 8 illustrates the performance of MKDE in a two dimensional setting.

In conclusion, we described MKDE, a Matérn-kernel-based, Bayesian, and non-parametric density estimator that can construct a smooth (probability) density function from an - even limited - set of data samples. The broad priors on the learned parameters, combined with the log-normal model and the Matérn kernel covariance structure, make MKDE very flexible and robust. Furthermore, the Bayesian inference framework allows for posterior uncertainty quantification for the reconstructed density. A software implementation of MKDE is available in NIFTy 7 and is also released as an open-source Python package (DENSe), which can be found at: https://ift.pages.mpcdf.de/public/dense/.

[6] For convenience we adopt the Fourier-transform convention in which the zeroth mode of the Fourier transform of a quantity corresponds to the integral of such quantity in position space. With this convention, we scale the proportionality factor which preserves the per-pixel standard deviation, denoted with a_s in Eq. 19 by $\sqrt{V}$, with V being the total volume of the position space.

[7] α is also scaled by $\sqrt{V}$ so that it does not depend on the volume in position space.

[8] We set the following priors on the signal parameters: $\alpha = [10^{-15}, 5.0]$, $a_s = (0.3 \pm 0.2)$, $k_0 = (4.0 \pm 3.0)$, and $\gamma_s = (-6.0 \pm 3.0)$, where the mean and standard deviation specify a Gaussian prior distribution for γ_s and log-normal distributions with the given mean and standard deviation for a_s and k_0.