Mathematical Modelling of Knowledge Acquisition and Consensus Formation in Populations

Ola Hössjer[1*], Daniel Andrés Díaz–Pachón[2] and J. Sunil Rao[3]

[1*]Department, Stockholm University, 106 91 Stockholm, Sweden.
[2]Division of Biostatistics, University of Miami, USA.
[3]Department of Biostatistics, University of Minnesota, USA.

*Corresponding author(s). E-mail(s): ola@math.su.se;
Contributing authors: DDiaz3@med.miami.edu; js-rao@umn.edu;

Abstract

In this article we study knowledge acquisition and consensus formation in populations, about a parameter $\boldsymbol{x}$ that belongs to opinion space $\boldsymbol{\mathcal{X}}$. We take a mixed Bayesian-frequentist approach, whereby agents in the populations form their beliefs individually from exponential family data as posterior distributions (the Bayesian component), with one element $\boldsymbol{x_0} \in \boldsymbol{\mathcal{X}}$ regarded as the true parameter value (a frequentist assumption). Consensus formation means that the agents' posterior distributions get increasingly similar, either as more data is collected, and/or as the agents iterate and update their beliefs among themselves. Knowledge acquisition additionally requires that the agents' posterior distributions get increasingly concentrated around $\boldsymbol{x_0}$. Within this framework, we provide general results for asymptotic knowledge acquisition and consensus formation to take place, for different data generating mechanisms, population structures and data missingness mechanisms (where the way data are discarded may reflect personal preferences). In particular, our approach makes it possible to model some agents as influencers, whereas others are more passive data collectors. A number of generalizations are also suggested, including the effect that some individuals generating fake data has on consensus formation and knowledge acquisition.

Keywords: Asymptotic results, Bayesian approach, consensus formation, exponential family, frequentist approach, knowledge acquisition, missing data, population structure

MSC Classification: 60F05 , 62A01 , 62C05 , 62C10 , 91D30

1 Introduction

How do humans acquire knowledge and/or reach consensus? As social beings we are dependent on one other, and not the least we learn from each other. This does not necessarily mean that we reach consensus, in particular when the proposition we acquire knowledge about is complex. The aim of this article is to present a matemathical model for knowledge acquisition (KA) and consensus formation (CF) in populations. The main motivation is that KA and CF should be modelled jointly, since CF without KA is not desirable for propositions that logically are either true or false. In the paper we will develop a framework that allows for a joint analysis of KA and CF, combining elements of Bayesian statistics [5] and frequentist statistics [27].

As a staring point of this work, we make use of the general theory of learning and KA for single individuals (or agents), presented in [21]. This theory has been applied to cosmology [12] and further developed in terms of maximum entropy-based statistical learning [11]. The underlying assumption of the theory is that learning corresponds to true beliefs, whereas KA requires more—justified, true beliefs [23]. This is formulated in terms of a statistical model with a parameter set $\mathcal{X}$ that corresponds to a set of possible opinions. The Bayesian component of the model is a posterior distribution on $\mathcal{X}$ that reflects the beliefs of an agent A. The frequentist component is the assumption that among all possible opinions in $\mathcal{X}$, one opinion x_0 is true whereas the others $x \neq x_0$ are not. Agent A learns and acquires knowledge about a proposition that is either true of false for each possible opinion $x \in \mathcal{X}$. More specifically, he learns about the proposition if he correctly finds out whether it is true or not, whereas KA additionally requires that he learns for the right reason, i.e., acquires belief in opinion x_0. In this paper we make the simplifying assumption that the proposition is true for all possible opinions $x \in \mathcal{X}$, so that learning always holds. Knowledge acquisition is then equivalent to increasingly accurate beliefs about x_0, as more data is collected, that is, a posterior distribution that gets increasingly concentrated around x_0.

In this article we will extend results of [21] for individuals to populations. These populations consist of a number of agents, each of which forms his beliefs through an individual-specific posterior distribution. It is assumed that the agents of the population are connected through a directed and weighted graph, where the directed weight from agent A_j to agent A_i quantifies how much A_j influences the belief formation of A_i. Within this framework, KA corresponds to the agents forming increasingly accurate beliefs about x_0 when more data is collected, whereas CF only requires that their posterior distributions get increasingly similar. The results of this article will contribute to a theoretical understanding of factors that explain KA and CF within populations. More specifically, our approach includes the following features:

1. The Bayesian-frequentist statistical framework for KA and CF in populations is very general. The statistical model allows us to analyze a variety of data types, corresponding to different exponential family statistical models.
2. Our model takes into account that individuals in the population receive different subsets of the total data set, according to a data missingness mechanism, where data is either missing completely at random, or in a biased way, to favor certain

opinions. Within this context we analyze how the data missingness mechanism affects KA and CF.

3. Since individuals communicate their beliefs to each other, our model makes it possible to investigate how different population structures (that is, the way in which individuals influence each other) affect KA and CF. This inlcudes scenarios when all individuals are equally influential, or those for which a few individuals influence the other members of the population.
4. We analyze dynamic aspects of KA and CF, when the size of the data set grows. In particular, we investigate under which conditions full KA and full CF is obtained asymptotically as the number of data points tends to infinity.
5. We study dynamic aspects of KA and CF, when individuals in the population iterate their beliefs among themselves, based on a given set of data. For instance, we investgate for which population structures full CF is attained when the number of iterations grows.
6. We assume that agents use conjugate priors for the given exponential family of data distributions. For this reason agents need not communicate their entire posterior distributions, but only a finite number of hyperparameters. For normally distributed data this is equivalent to communicating the current best guess of x_0 and the aposteriori uncertainty of this guess. In particular we demonstrate that higher certainty leads to higher influence among the other agents.

Our framework for belief formation in a population differs from an Exponential family random graph model [17, 30] in that the weighted graph of individuals and their mutual influences is fixed, not random. In our model it is rather the data that the population is exposed to that is random and drawn from an exponential family. Our model is more reminiscent of a Bayesian network [34], with the difference that i) directed edges not only signify causal relationships, but the weights attached to these edges quantify the strengths of these causal relationships, and ii) we use a Bayesian-frequentist approach, with data generated from a supposedly true model with parameter value x_0.

Previous mathematical approaches to consensus formation can broadly be divided into Bayesian and non-Bayesian models [3, 13, 24, 33]. The most well known non-Bayesian model is the the DeGroot model ([4, 8, 9, 25]), where opinions of all agents are treated as fixed numbers in opinion space and updated recursively in discrete time, based on input from other agents. Continuous time extensions of this model are frequently used in computer science and control theory to model consensus formation [37]. Another non-Bayesian model of consensus formation is the Deffuant model ([5, 10, 19, 32]; see also [7, 18] for overviews) where opinions are treated as random variables on opinion space, and updated in continuous time based on inputs form one single agent at a time. Another, somewhat related continuos time non-Bayesian model also has pairwise encounters between agents, with so called forceful agents being more influential than the remaining regular agents [1]. Some of these non-Bayesian models are also related to interacting particle systems on random graphs [26, 28], with interactions interpreted as social influences.

Since data is not explicitly part of non-Bayesian models, they are typically regarded as simplified version of Bayesian models. The Bayesian framework is indeed more

involved, since agents form their opinions as posterior distributions (rather than atoms) on the opinion space, based priors and likelihoods, before they communicate these posterior distributions to other individuals of the population, which in turn leads to updated posterior distributions. However, our Bayesian, discrete-time approach leads to surprisingly simple postereior distribution, for two reasons: First, we assume that agents update their posteriors proportionally to the product of other agents' posteriors, raised to the powers of their influences (so called logarithmic pooling, see for instance [14]). This implies that the impact of these other agents is very explicit, similarly as for the non-Bayesian DeGroot model. Second, since we make use of conjugate priors, the resulting Bayesian consensus formation algorithm only depends on hyperparameters of the prior distribution, which are updated when agents receive data and opinions from other agents. Another novelty of our Bayesian approach is that we additionally make a frequentist assumption that one parameter value (opinion) x_0 is true. Although some articles consider Bayesian social learning within a frequentist context (see for instance [2]), to the best of our knowledge, there is no previous systematic attempt to treat KA and CF jointly for a wide range of data types and data distributions. This is important, since consensus formation is not an isolated goal of itself for opinions about parameters that have a correct value x_0. If a consensus opinion emerges that turns out to be false, there is CF but no KA. A somewhat related Bayesian approach to consensus formation, for sensor networks, it is to assume that agents receive different datasets, and require that the agents eventually reach consensus about an estimate $\hat{x}$ of x_0, based on data from all agents [20, 38]. The closer to x_0 this estimate is, the higher is the degree of KA of the network.

This article is organized as follows: In Section 2 we introduce the model of [21] for KA of individuals. Then in Section 3 we define a model for KA and CF in populations, where the posterior distributions of all agents depend on their prior distributions; the likelihood function of data; the directed, weighted graph of influences, and the data missingness mechanisms. Section 4 provides asymptotic results for KA and CF, as the number of data points grows, whereas in Section 5 we consider KA and CF when agents iterate their beliefs among themselves. The framework of KA and CF is exemplified in Section 6 when data either follow a Bernoulli or a normal distributon. A discussion is provided in Section 7, and mathematical proofs are gathered in Appendix A.

2 Model for single individuals

2.1 Statistical model

Suppose one individual (or agent) A has access to data

$$\boldsymbol{D} = (D_1, \ldots, D_N) \tag{1}$$

of size N. It is assumed that the data points D_k are independent and identically distributed random variables $D_k \in \mathcal{D}$ with a density $f(\cdot; x)$ that is indexed by a parameter $x \in \mathcal{X}$, whose true value x_0 is unknown. The prior distributon P_0 on $\mathcal{X}$ summarizes the agents's prior beliefs about x_0. Then, after seeing data $\boldsymbol{D}$, A updates

his prior beliefs to a posterior distribution

$$P(x) = P(x|\boldsymbol{D}) = \frac{P_0(x)L(\boldsymbol{D}|x)}{L(\boldsymbol{D})} \tag{2}$$

according to Bayes' rule. Here $L(\boldsymbol{D}|x)$ is the likelihood function of agent A, whereas $L(\boldsymbol{D}) = \int_{\mathcal{X}} P_0(x)L(\boldsymbol{D}|x)dx$ is obtained by integrating the likelihood over the parameter space $\mathcal{X}$ with respect to the agent's prior beliefs. If A correctly specifies the likelihood function, it is given by

$$L(\boldsymbol{D}|x) = \prod_{k=1}^{N} f(D_k; x). \tag{3}$$

Note that this framework combines elements of frequentism (the assumption that one parameter value x_0 is true) and Bayesianism (modeling degrees of beliefs with prior and posterior distributions). This is a key aspect of our model that later on will be used for developing the concepts of KA and CF for groups of individuals.

The parameter space $\mathcal{X}$ is either discrete (a countable set) or a subset of q-dimensional Euclidean space $\mathbb{R}^q$. In this article we will mainly focus on the latter. Each data point is generated from an exponential family (cf. Section 1.5 of [27])

$$f(d; x) = \exp\left[\sum_{r=1}^{q} \eta_r(x)T_r(d) - B(x)\right] h(d), \tag{4}$$

where $\{\eta_r(x)\}_{r=1}^q$ are the natural parameters and $\{T_r(d)\}_{r=1}^q$ the real-valued sufficient statistics. We will assume a conjugate prior

$$P_0(x) = P(x; \alpha) \propto \exp\left[\sum_{r=1}^{q} \eta_r(x)\alpha_r - \sum_{r=1}^{s} \alpha_{q+r}B_r(x)\right] \tag{5}$$

for agent A, for some integer $s \geq 0$ and functions $\{B_r(x)\}_{r=1}^s$ that satisfy $\sum_{r=1}^s B_r(x) = B(x)$. The hyperparameter vector of the prior distribution is $\alpha = (\alpha_1, \ldots, \alpha_{q+s})$. Inserting (4)–(5) into (2)–(3) we find that the posterior distribution is given by

$$P(x) = P(x; \bar{\alpha}), \tag{6}$$

with a data-dependent hyperparameter vector

$$\bar{\alpha} = (\bar{\alpha}_1, \ldots, \bar{\alpha}_{q+s}) = \alpha + \sum_{k=1}^{N} (T_1(D_k), \ldots, T_q(D_k), 1, \ldots, 1). \tag{7}$$

Example 1 (Bernoulli distribution.) A Bernoulli distribution $D \sim \mathrm{Be}(x)$ has probability function

$$f(d; x) = x^d(1-x)^{1-d} = \exp\left[d \log \frac{x}{1-x} + \log(1-x)\right]$$

for $0 \leq x \leq 1$. This corresponds to an exponential family (4) with $q = 1$, $T(d) = d$, $\eta(x) = \log[x/(1-x)]$ and $B(x) = -\log(1-x)$. A beta prior $X \sim B(a, b)$ has a probability density function

$$P_0(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \propto \exp\left[(a-1)\log\frac{x}{1-x} + (a+b)\log(1-x)\right].$$

It is a conjugate prior (5) with $s = 1$, $\alpha_1 = a-1$ and $\alpha_2 = a+b$. The posterior distribution (6)-(7) corresponds to a beta distribution $\mathrm{Be}(\bar{a}, \bar{b})$, with $\bar{\alpha}_1 = \bar{a} - 1$, $\bar{\alpha}_2 = \bar{a} + \bar{b}$, $\bar{a} = a + \sum_{k=1}^{N} D_k$ and $\bar{b} = b + N - \sum_{k=1}^{N} D_k$. □

Example 2 (Exponential distribution.) An exponential distribution $D \sim \mathrm{Exp}(x)$ with intensity parameter $x > 0$ has density function

$$f(d; x) = xe^{-xd} = \exp(-dx + \log x),$$

so that $q = 1$, $T(d) = d$, $\eta(x) = -x$, and $B(x) = -\log x$ in (5). A gamma conjugate prior $P_0 \sim \Gamma(a, b)$ has a density function

$$P_0(x) = \frac{1}{\Gamma(b)} a^b x^{b-1} e^{-ax} \propto \exp[-ax + (b-1)\log x]$$

that corresponds to $s = 1$, $\alpha_1 = a$ and $\alpha_2 = b - 1$ in (5). The posterior distribution (6)-(7) is another gamma distribution $\Gamma(\bar{a}, \bar{b})$ with $\bar{\alpha}_1 = \bar{a}$, $\bar{\alpha}_2 = \bar{b} - 1$, $\bar{a} = a + \sum_{k=1}^{N} D_k$ and $\bar{b} = b + N$. □

Example 3 (Normal distribution with known variance.) A normal distribution $D \sim N(x, \sigma^2)$ with mean $-\infty < x < \infty$ and known variance $\sigma^2 = 1/\kappa$ has density function

$$f(d; x) = \sqrt{\frac{\kappa}{2\pi}} \exp\left[-\frac{\kappa}{2}(d-x)^2\right] \propto \exp\left(\kappa x d - \frac{\kappa x^2}{2}\right),$$

with $q = 1$, $T(d) = d$, $\eta(x) = \kappa x$, and $B(x) = \kappa x^2/2$ in (4). A conjugate prior $X \sim N(m, c^{-1})$ has a density function

$$P_0(x) = \sqrt{\frac{c}{2\pi}} \exp\left[-\frac{c}{2}(x-m)^2\right] \propto \exp\left(cmx - \frac{cx^2}{2}\right)$$

that corresponds to $s = 1$, $\alpha_1 = cm\kappa^{-1}$ and $\alpha_2 = c\kappa^{-1}$ in (5). The posterior distribution (6)-(7) is normal $N(\bar{m}, \bar{c}^{-1})$, with $\bar{\alpha}_1 = \bar{c}\bar{m}\kappa^{-1}$, $\bar{\alpha}_2 = \bar{c}\kappa^{-1}$, $\bar{m} = (cm + \kappa N\bar{D})/\bar{c}$, $\bar{c} = c + \kappa N$ and $\bar{D} = \sum_k D_k/N$. In particular, the posterior mean $\bar{m}$ is a weighted average of m and $\bar{D}$, with a larger weight attached to $\bar{D}$ the larger the dataset is. □

Example 4 (Normal distribution with unknown variance.) The normal distribution $D \sim N(x_1, x_2^{-1})$ with mean $-\infty < x_1 < \infty$ and inverse variance $x_2 > 0$ has a density function

$$f(d; x) = \sqrt{\frac{x_2}{2\pi}} \exp\left[-\frac{x_2}{2}(d-x_1)^2\right] \propto \exp\left[x_1 x_2 d - \frac{x_2}{2}d^2 - \frac{x_1^2 x_2}{2} + \frac{\log(x_2)}{2}\right],$$

with $q = 2$ and $x = (x_1, x_2)$. This corresponds to $T_1(d) = d$, $T_2(d) = d^2$, $\eta_1(x) = x_1 x_2$, $\eta_2(x) = -x_2/2$, and $B(x) = x_1^2 x_2/2 - \log(x_2)/2$ in (4). The conjugate prior is $X_2 \sim \Gamma(a, b)$ and $X_1 | X_2 = x_2 \sim N(m, (cx_2)^{-1})$. This gives a prior density

$$\begin{aligned} P_0(x) &= \tfrac{a^b}{\Gamma(b)} x_2^{b-1} e^{-ax_2} \cdot \sqrt{\tfrac{cx_2}{2\pi}} e^{-\frac{cx_2}{2}(x_1-m)^2} \\ &\propto \exp\left[cmx_1x_2 - (a + \tfrac{cm^2}{2})x_2 - \tfrac{cx_1^2x_2}{2} + (b - \tfrac{1}{2})\log(x_2))\right], \end{aligned}$$

that corresponds to $s = 2$ in (5), with $\alpha_1 = cm$, $\alpha_2 = 2a + cm^2$, $\alpha_3 = c$, $\alpha_4 = 2b - 1$, $B_1(x) = x_1^2 x_2/2$, and $B_2(x) = -\log(x_2)/2$. The posterior distribution (6)-(7) simplifies to $X_2 \sim \Gamma(\bar{a}, \bar{b})$ and $X_1|X_2 = x_2 \sim N(\bar{m}, (\bar{c}x_2)^{-1})$, where $\bar{\alpha}_1 = \bar{c}\bar{m} = cm + \sum_k D_k$, $\bar{\alpha}_2 = 2\bar{a} + \bar{c}\bar{m}^2 = 2a + cm^2 + \sum_k D_k^2$, $\bar{\alpha}_3 = \bar{c} = c + N$, and $\bar{\alpha}_4 = 2\bar{b} - 1 = 2b - 1 + N$. The parameters of the posterior can be expressed as $\bar{c} = c + N$, $\bar{b} = b + N/2$, $\bar{m} = (cm + N\bar{D})/(c + N)$, $\bar{D} = \sum_k D_k/N$, and

$$\begin{aligned} \bar{a} &= a + \tfrac{1}{2}cm^2 + \tfrac{1}{2}\textstyle\sum_k D_k^2 - \frac{c+N}{2}(\frac{c}{c+N}m + \frac{N}{c+N}\bar{D})^2 \\ &= a + \tfrac{cN}{2(c+N)}(\bar{D} - m)^2 + \tfrac{1}{2}\textstyle\sum_k (D_k - \bar{D})^2. \end{aligned}$$

□

2.2 Knowledge acquisition

Loosely speaking, an agent acquires knowledge about x_0 when the posterior distribution P is more concentrated around x_0 than the prior P_0. This can be formalized by introducing a metric $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ on $\mathcal{X}$. Let also $B(x_0, \varepsilon) = \{x \in \mathcal{X}; d(x, x_0) < \varepsilon\}$ refer to the open ball of radius $\varepsilon > 0$ around x_0.

Definition 1 (Knowledge acquisition of one individual.) We say that agent A has acquired knowledge about x_0 if

$$\begin{aligned} &P[B(x_0, \varepsilon)] \geq P_0[B(x_0, \varepsilon)], \text{ for all } \varepsilon > 0, \\ &P[B(x_0, \varepsilon)] > P_0[B(x_0, \varepsilon)], \text{ for at least one } \varepsilon > 0. \end{aligned} \tag{8}$$

□

Note that (8) is equivalent to $d(X, x_0)$ being stochastically smaller than $d(X_0, x_0)$, if $X \sim P$ and $X_0 \sim P_0$ are two random variables on $\mathcal{X}$ distributed according to P and P_0 respectively. The abovementioned definition (8) of KA is slightly weaker than Definition 2 of [21].

2.3 Asymptotic knowledge acquisition

We can make use of asymptotic Bayesian theory [15] to formulate how an agent increases his knowledge when the number of data points grows ($N \to \infty$). The following definition specifies when an agent acquires full knowledge asymptotically, at given rate.

Definition 2 (Asymptotic knowledge acquisition of one individual as the number of data points increases.) An agent A acquires full knowledge asymptotically about x_0 as $N \to \infty$ at rate $\varepsilon_N \to 0$ if the posterior distribution P in (2) converges in probability towards a one point distribution δ_{x_0} at rate ε_N. That is,

$$P[B(x_0, a_N \varepsilon_N)^c] \to 0 \tag{9}$$

as $N \to \infty$ for any sequence $a_N \to \infty$, where $B(x_0; \varepsilon)^c = \mathcal{X} \setminus B(x_0; \varepsilon)$. □

Asymptotic KA often holds for an agent with a correctly specified likelihood (3). When the parameter space $\mathcal{X} \subset \mathbb{R}^q$ is a convex and open subset of q-dimensional

Euclidean space, it is possible to make use of large sample, asymptotic theory and find the convergence rate at which P approaches δ_{x_0}, a point mass at x_0. Under mild regularity conditions it turns out that the posterior distribution approaches x_0 at rate $\varepsilon_N = 1/\sqrt{N}$. More specifically, it was proved in [21] that if $X = X_N \sim P$ is distributed according to the posterior based on N data points, we have weak convergence

$$\sqrt{N}(X_N - x_0) \xrightarrow{\mathcal{L}} N(0, 2J(x_0)^{-1}) \tag{10}$$

towards a multivariate normal distribution as $N \to \infty$, with asymptotic covariance matrix twice the inverse of the Fisher information matrix

$$J(x) = \int_{\mathcal{D}} \psi(z;x)^T \psi(z;x) f(z;x) dz, \tag{11}$$

evaluated at x_0. Here

$$\psi(d;x) = \frac{f'(d;x)}{f(d;x)} = \sum_{r=1}^{q} \eta_r'(x) T_r(d) - B'(x) \tag{12}$$

is the q-dimensional likelihood score function, $f'(d;x)$ refers to the gradient vecor of $f(d;x)$ when differentiating with respect to x, whereas T in (11) corresponds to vector transposition. The asymptotic result (10) makes use of asymptotic theory of maximum likelihood estimates and the Bernstein-von Mises theorem for the asymptotic behaviour of posterior distributions.

3 Model for populations

3.1 Statistical model

3.1.1 Population structure and choice of influence weights

A population of n agents is modelled as a network [31]. More specifically, the population is a weighted, bi-directed random graph $\mathcal{G} = (\mathcal{A}, \boldsymbol{W})$, whose vertex set $\mathcal{A} = (A_1, \cdots, A_n)$ corresponds to the n agents, whereas

$$\boldsymbol{W} = (w_{ji})_{j,i=1}^{n} \tag{13}$$

is a square matrix containing the weights between all n^2 ordered pairs of vertices. The weight $w_{ji} \geq 0$ assigned to the directed edge (A_j, A_i) quantifies the influence that A_j has on A_i in terms of KA. These weights need to be normalized in some way. One option is to assume that each agent sets his own influence

$$w_{ii} = 1 \tag{14}$$

to unity, whereas the influences $\{w_{ji}; 1 \leq j \leq n, j \neq i\}$ of the other agents on A_i can be smaller or larger than 1. Another option is to assume that that

$$\sum_{j=1}^{n} w_{ji} = 1 \tag{15}$$

for $i = 1, \ldots, n$, so that w_{ji} is the relative influence that A_j has on A_i. In this setting, the self-influence weight w_{ii} is a measure of agent's A_i assimilation of others' opinions: If $w_{ii} = 1$ A_i is called *stubborn*, whereas $w_{ii} = 0$ makes A_i completely open to be influenced by others [36]. Observe that if more than one agent is stubborn, CF is typically impossible, unless the stubborn agents receive the same data and interpret it in the same way. For both types (14) and (15) of normalization, it is also possible to let the weights w_{ji} depend on data. Such a model is however more complicated to analyze. Regardless of whether weights depend on data or not, one may renormalize the n weights $w_{1i}, \ldots, w_{ni}$ towards A_i multiplicatively according to

$$w_{ji} \mapsto a_i w_{ji}, \quad j = 1, \ldots, n, \tag{16}$$

for some constant a_i. For instance, it is possible to re-normalize weights according to (16), with

$$a_i = 1/\sum_{j=1}^{n} w_{ji}, \tag{17}$$

for $i = 1, \ldots, n$, so that (15) holds. The population structure is determined by the collection $\boldsymbol{W} = (w_{ji})$ of weights of all directed edges. Below we give four examples of population structures with weight normalized according to (14). In all these four examples, a renormalization (17) will change the weights in order to satisfy (15).

Example 5 (Symmetric population.) The first type of population structure is symmetric, corresponding to

$$w_{ji} = 1(i = j) + \frac{w}{n-1} 1(i \neq j), \tag{18}$$

for some $w \geq 0$. Equation (18) implies that each agent A_i makes decisions in isolation when $w = 0$, whereas A_i is influenced (to the same degree) by all other agents when $w > 0$, with $\mathcal{G}$ a complete graph. The parameter w quantifies the total influence of all other agents in comparison to A_i's own assessment of data. When $w > 1$, the other agents have a larger total influence on A_i than A_i's own assessment. □

Example 6 (Star-shaped population.) The second star-shaped type of population has

$$w_{ji} = 1(i = j) + w1(j = 1, i \neq 1). \tag{19}$$

This population has one influencer A_1 when $w > 0$, whereas all other agents are followers. The larger $w > 0$ is, the more influential A_1 is. When $w > 1$, A_1 has a larger influence on all other agents compared to their own assessment of data. As long as $w > 0$, $\mathcal{G}$ is cone graph with A_1 the special vertex that connects to all other vertices. As w gets large, the $1(i = j)$ term of (19) will have no impact and $\mathcal{G}$ becomes isomorphic to a tree, with A_1 determining the opinons of all other agents. □

Example 7 (Linearly ordered population.) A linearly ordered population has

$$w_{ji} = 1(i = j) + w1(i = j + 1),$$

so that A_i influences A_{i+1} for $i = 1, \dots, n-1$, to an extent determined by $w > 0$. □

Example 8 (Subdivided population.) Suppose a population of size $n = mr$ is divided into m subpopulations of equal size r. Then order individuals so that those agents A_i for which $i \in \{(t-1)r + 1, \dots, tr\}$ belong to Subpopulation t, for $t = 1, \dots, m$. Let

$$w_{ji} = 1(i = j) + w_1 1(i \neq j, [(i-1)/r] = [(j-1)/r]) + w_2 1([(i-1)/r] \neq [(j-1)/r]),$$

where $w_2 \geq 0$ and $w_1 > w_2$ quantify strength of influence between and within subpopulations respectively. □

3.1.2 Data collection and missing data

Suppose A_i has access to a subset $\boldsymbol{D}_i = \{D_k : M_{ik} = 1\} \subset \boldsymbol{D}$ of the random data vector (1), where M_{ik} is a binary random variable that determines whether A_i misses the random data point D_k ($M_{ik} = 0$) or not ($M_{ik} = 1$). We will assume that data of A_i are missing at random (MAR), see [22, 29]. By this we mean that the probability of a data point d not being missed is an agent-specific function

$$v_i(d) = P(M_{ik} = 1 | D_k = d) \tag{20}$$

of d itself, and independent of k. It is further assumed that various random data points D_k are missed independently, in the sense that

$$(D_k, \boldsymbol{M}_k) \text{ are independent for } k = 1, \dots, N, \tag{21}$$

where $\boldsymbol{M}_k = (M_{1k}, \dots, M_{nk})$ is a binary vector of missingness indicators of all agents for data point D_k. Additionally, we require

$$\{M_{ik}\}_{i=1}^n | D_k \text{ are conditionally independent} \tag{22}$$

for $k = 1, \dots, N$, so that no other joint factors than data influence the missingness mechanism of the agents. The data missingness mechanism (20) will have a large impact on whether KA and CF is attained asymptotically. Below we present two different MAR missingness mechanisms:

Example 9 (Data missing completely at random.) The simplest missingness mechanism is obtained if data of each agent A_i is missing completely at random (MCAR, see [29]), according to

$$v_i(d) = v_i \tag{23}$$

for some $0 \leq v_i \leq 1$. It is reasonable then that KA is achieved, and that cooperation between agents (large weights w_{ji} for $i \neq j$) is beneficial in order to attain knowledge faster as more data is collected. □

Example 10 (Biased missing data.) For a a missingness mechanism such that data consistent with x_0 is discarded more frequently, it is less likely that KA is attained. An example of such a missingness mechanism for agent A_i is

$$v_i(d) = v_i^{\max} \left(\frac{f(d; y_i)}{f_\infty(y_i)} \right)^{s_i}, \tag{24}$$

where $0 \leq v_i^{\max} \leq 1$, $f(d; x)$ is the likelihood in (4) of a data point d, $y_i \in \mathcal{X}$ is a preferred parameter value of A_i, $f_\infty(y_i) = \sup_{d \in \mathcal{D}} f(d; y_i)$, whereas $s_i \geq 0$ determines how much A_i prefers y_i. Note that $v_i(d)$ is defined in such a way that $0 \leq v_i(d) \leq v_i^{\max} \leq 1$ is guaranteed. For a fixed $v_i^{\max}$, the larger s_i is, the more data A_i misses, with $s_i = 0$ corresponding to data MCAR, with a proportion $v_i = v_i^{\max}$ of data retained. Intuitively, we expect that KA will not be attained if the population has some very influential agent A_i (for instance $i = 1$ in (19), with w large) whose missingness mechanism satisifies (24), with $v_i^{\max} = 1$, s_i large and y_i far away from x_0. □

3.1.3 Likelihoods and posteriors

The likelihood of agent A_i will involve the likelihoods $f(D_k; x)$ in (4) of all random data points D_k that he receives ($M_{ik} = 1$) and possibly also a correction for missing data. In order to handle missing data, we will assume that A_i believes the missingness mechanism is determined by the function $u_i : \mathcal{D} \times \mathcal{X} \to [0, 1]$, where $u_i(d; x) f(d; x)$ is the agent's estimate of $P(M_{ik} = 1, D_k = d)$. Possible choices of u_i include

$$u_i(d; x) \equiv 1 \tag{25}$$

and

$$u_i(d; x) = v_i(d) \left(\int_{\mathcal{D}} v_i(z) f(z; x) dz \right)^{-1} \propto \left(\int_{\mathcal{D}} v_i(z) f(z; x) dz \right)^{-1}. \tag{26}$$

In the former case, missing data are not at all corrected for. In the latter case, missing data are fully corrected for and the proportionality of (26) is between quantities that differ by $v_i(d)$, which here is a multiplicative constant not depending on x. As we will find below, such constants have no impact on the posterior distribution of A_i. Incorporating the fact that some data are missing, and a missing data correction, we obtain a likelihood function

$$L_i(\boldsymbol{D}_i | x) = \prod_{k; M_{ik} = 1} [u_i(D_k; x) f(D_k; x)] = \prod_{k=1}^{N} [u_i(D_k; x) f(D_k; x)]^{M_{ik}} \tag{27}$$

of agent A_i, which is correctly specified only if (26) holds. The posterior distribution of A_i is formed in two steps. In the first step of belief formation, A_i obtains a posterior distribution

$$P_{i1}(x) = P_{i1}(x | \boldsymbol{D}_i) = \frac{P_{i0}(x) L_i(\boldsymbol{D}_i | x)}{L_i(\boldsymbol{D}_i)} \tag{28}$$

based on his own prior distribution P_{i0} and likelihood function $L_i(\boldsymbol{D}_i | x)$. When $\mathcal{X} \subset \mathbb{R}^q$ is a subset of a Euclidean space and the data points are generated from an

exponential family (4), we will assume a conjugate prior

$$P_{i0}(x) = P(x;\alpha_i) \propto \exp\left[\sum_{r=1}^{q} \eta_r(x)\alpha_{ir} - \sum_{r=1}^{s} \alpha_{i,q+r} B_r(x)\right] \tag{29}$$

for agent A_i, with hyperparameter vector $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{i,q+s})$. Typically, α_i is chosen in such a way that $P_{i0}(\cdot)$ is concentrated around some apriori preferred parameter value $y_i \in \mathcal{X}$ of A_i.
In the second and final step of belief formation, A_i updates his posterior distribution based on inputs from the other agents to $P_i(x) = P_{i2}(x)$, as

$$P_i(x) \propto \prod_{j=1}^{n} P_{j1}(x)^{w_{ji}}. \tag{30}$$

Combining (27)-(30), we find that

$$\begin{aligned} P_i(x) &\propto \textstyle\prod_{j=1}^{n} P_{j0}(x)^{w_{ji}} \cdot \prod_{k=1}^{N} \prod_{j=1}^{n} [u_j(D_k;x) f(D_k;x)]^{w_{ji} M_{jk}} \\ &= \textstyle\prod_{j=1}^{n} P_{j0}(x)^{w_{ji}} \prod_{k=1}^{N} [\bar{u}_{ik}(D_k;x) f(D_k;x)]^{\bar{M}_{ik}}, \end{aligned} \tag{31}$$

where

$$\bar{M}_{ik} = \sum_{j=1}^{n} w_{ji} M_{jk} \tag{32}$$

corresponds to how much data point D_k influences the posterior of A_i, whereas

$$\bar{u}_{ik}(D_k;x) = \left[\prod_{j=1}^{n} u_j(D_k;x)^{w_{ji} M_{jk}}\right]^{1/\bar{M}_{ik}} \tag{33}$$

quantifies how A_i adjusts for the sampling of D_k, when inputs from the other agents are taken into account.

Remark 1 (The effect of renormalized weights on posteriors.) If the influence weights $w_{1i}, \ldots, w_{ni}$ of agent A_i are normalized as (14), the updated posterior P_i of A_i in (30)–(31) is proportional to a weighted geometric average of all agents' posteriors. Note also that the multiplicative renormalization (16) of the n weights $w_{1i}, \ldots, w_{ni}$ towards A_i changes A_i's posterior in (30)–(31) from $P_i(x)$ to $P_i(x)^{a_i}/Z_i$, where Z_i is a normalizing constant assuring that the modified posterior still integrates to 1. The maximum aposteriori (MAP) estimate $\hat{x}_i = \arg\max_x P_i(x)$ of agent A_i is on one hand unaffected by the renormalization in (16). On the other hand, the larger a_i is, the more concentrated is A_i's posterior around $\hat{x}_i$. This implies, for instance, that P_i is more concentrated around $\hat{x}_i$ when (14) holds, compared to (15). We will refer to $\sum_j w_{ji}$ as the boldness of A_i. The effect of normalizing the influence weights according to (16) is to multiply A_i's boldness with a_i, so that a higher boldness leads to a higher degree of confidence in $\hat{x}_i$. □

If missing data is fully corrected for, according to (26), then

$$\bar{u}_{ik}(d;x) \propto \left[\prod_{j=1}^{n}\left(\int_{\mathcal{D}} v_i(z)f(z;x)dz\right)^{-w_{ji}M_{jk}}\right]^{1/\bar{M}_{ik}},$$

whereas if missing data are not at all corrected for, according to (25), then

$$\bar{u}_{ik}(d;x) = 1. \tag{34}$$

We will mostly assume that missing data are not corrected for, so that (25) and (34) hold. The missingness mechanism (20) will then have a large impact on whether agent A_i acquires knowledge or not (whereas knowledge *is* acquired for most missingness mechanisms when missing data are corrected for, according to (26)). Generalizing (7), we find that the posterior distributions of agent A_i in (28) and (30), simplify to

$$P_{i1}(x) = P_{i1}(x;\bar{\alpha}_{i1})$$

and

$$P_i(x) = P(x;\bar{\alpha}_i) \propto \prod_{j=1}^{n} P_{j0}(x)^{w_{ji}} \prod_{k=1}^{N} f(D_k;x)^{\bar{M}_{ik}}, \tag{35}$$

respectively when missing data are not corrected for, as in (34), with

$$\bar{\alpha}_{i1} = \alpha_i + \sum_{k=1}^{N} M_{ik}(T_1(D_k),\ldots,T_q(D_k),1,\ldots,1)$$

and

$$\bar{\alpha}_i = \bar{\alpha}_{i2} = \sum_{j=1}^{n} w_{ji}\alpha_j + \sum_{k=1}^{N} \bar{M}_{ik}(T_1(D_k),\ldots,T_q(D_k),1,\ldots,1) \tag{36}$$

the posterior hyperparameters of A_i after the first and second steps of belief formation respectively. Note in particular that it suffices in Step 2 for agents to communicate their $(q+s)$-dimensional hyperparameters $\bar{\alpha}_{i1}$ to those other agents that they influence (or another $(q+s)$-dimensional vector isomorphic to $\bar{\alpha}_{i1}$). For instance, for normal data (Example 3) with first step posteriors $P_{i1} \sim N(\hat{x}_{i1}, \sigma_{i1}^2)$, it suffices for agents to communicate their MAP estimates $\hat{x}_{i1}$ of x_0, and their aposteriori standard deviations σ_{i1}, to each other, in order for them to compute their updated posteriors P_{i2}.

3.2 Knowledge acquisition for populations

In this section we extend the notions of KA and asymptotic KA for individuals (Definitions 1-2) to populations. To this end, let

$$\begin{array}{rl} P_0 &= \sum_{i=1}^{n} P_{i0}/n, \\ P &= \sum_{i=1}^{n} P_i/n, \end{array} \tag{37}$$

be the average prior and average posterior distribution in the whole population.

Definition 3 (Knowledge acquisition in a population.) We say that agent A_i of a population of size n has acquired knowledge about x_0 if his prior distribution P_{i0} and posterior distribution P_i in (30) satisfy

$$\begin{aligned} P_i[B(x_0,\varepsilon)] &\geq P_{i0}[B(x_0,\varepsilon)], \text{ for all } \varepsilon > 0, \\ P_i[B(x_0,\varepsilon)] &> P_{i0}[B(x_0,\varepsilon)], \text{ for at least one } \varepsilon > 0. \end{aligned} \tag{38}$$

If (38) holds, with P_0 and P in (37) in place of P_{i0} and P_i, the population collectively has acquired knowledge about x_0. □

Definition 4 (Asymptotic full knowledge acquisition in a population as the number of data points increases.) We say that agent A_i of a population of size n acquires full knowledge asymptotically about x_0 as $N \to \infty$ at rate $\varepsilon_N \to 0$ if his posterior distribution P_i in (30) converges in probability towards a one point distribution δ_{x_0} at rate ε_N. That is,

$$P_i[B(x_0, a_N\varepsilon_N)^c] \to 0 \tag{39}$$

as $N \to \infty$ for any sequence $a_N \to \infty$. Analogously, the population acquires full knowledge asymptoticallly at rate $\varepsilon_N \to 0$, if (39) holds with P in (37) in place of P_i. □

3.3 Consensus formation

The degree of consensus between the n agents in $\mathcal{A}$ corresponds to how much their posterior distributions $P_1, \ldots, P_n$ in (37) align. In order to make this more precise, we introduce a distance $\delta(\cdot,\cdot)$ between probability distributions on the parameter space $\mathcal{X}$. The distance between two distributions is non-negative, and it equals zero only if the two distributions are identical. With P as in (37), let

$$\delta_i = \delta(P_i, P) \tag{40}$$

be the distance between the posterior of A_i and that of the whole population. Depending on the form of $\mathcal{X}$, the average distance to the population's posterior, for all agents, is chosen as one of

$$\delta = \begin{cases} \sum_{i=1}^n \delta_i/n, \\ (\sum_{i=1}^n \delta_i^2/n)^{1/2}. \end{cases} \tag{41}$$

The degree of consensus in the population is then defined as

$$C = \left(1 - \frac{\delta}{\delta_{\max}}\right)_+, \tag{42}$$

where $\delta_{\max}$ is an appropriately chosen upper bound on δ, and $z_+ = \max(z, 0)$ is the positive part of z. Consequently, C is a number between 0 and 1, with a larger value representing more consensus, and with $C = 1$ corresponding to full consensus.

Example 11 (Discrete parameter spaces) Suppose $\mathcal{X}$ is a countable set. It is possible then to use the Kullback-Leibler divergence

$$\delta(Q_1, Q_2) = d_{KL}(Q_1||Q_2) = \sum_{x\in\mathcal{X}} Q_1(x) \log \frac{Q_1(x)}{Q_2(x)}$$

as a distance measure between probability distributions Q_1 and Q_2 on $\mathcal{X}$. It follows from (37) and (40) that $P_i(x)/P(x) \leq n$ and $0 \leq \delta_i \leq \log(n)$, so that $\delta_{\max} = \log(n)$ can be used in the consensus definition (42). A second possibility is to use the Hellinger distance

$$\delta(Q_1, Q_2) = \sqrt{\frac{1}{2} \sum_{x\in\mathcal{X}} (\sqrt{Q_2(x)} - \sqrt{Q_1(x)})^2}.$$

The Hellinger distance is a metric, since it is symmetric and satisifies the triangle inequality. It is also bounded, $0 \leq \delta(Q_1, Q_2) \leq 1$, and it is therefore appropriate to put $\delta_{\max} = 1$ in (42). A third possibility is to use a quasi metric, i.e. a distance that satisfies the triangle inequality but not necessarily is symmetric [39]. This is of interest since the arguments of (40) correspond to beliefs of A_i and the whole population, respectively, and the distance of this equation quantifies how outlying A_i's posterior is in relation to the whole population. □

Example 12 (Euclidean parameter spaces) Suppose $\mathcal{X}$ is a connected and convex subset of $\mathbb{R}^q$, with $d(\cdot,\cdot)$ the q-dimensional Euclidean distance. Let Σ be a positive definite covariance matrix of order q that normalizes distances between elements of $\mathcal{X}$. For two distributions Q_1 and Q_2 on $\mathcal{X}$, put

$$\begin{aligned} \delta^2(Q_1, Q_2) &= (E_{Q_2}(X) - E_{Q_1}(X))\Sigma^{-1}(E_{Q_2}(X) - E_{Q_1}(X))^T \\ &= d(\Sigma^{-1/2}E_{Q_1}(X), \Sigma^{-1/2}E_{Q_2}(X)). \end{aligned} \tag{43}$$

Then (40) reduces to

$$\delta_i = \left[(m_i - m)\Sigma^{-1}(m_i - m)^T\right]^{1/2}, \tag{44}$$

where

$$\begin{aligned} m_i &= E_{P_i}(X), \\ m &= E_P(X) \end{aligned} \tag{45}$$

is the expected value of the posterior of A_i and the whole population respectively. If also $\mathcal{X}$ is bounded, it is convenient to let $\delta_{\max} = \max\{d(\Sigma^{-1/2}x, \Sigma^{-1/2}y);\, x, y \in \mathcal{X}\}$ be the diameter of $\Sigma^{-1/2}\mathcal{X}$ in the definition of the consensus variable (42). In particular, if $\Sigma = I_q$ is the identity matrix of order q, it follows that $\delta(Q_1, Q_2) = d(E_{Q_1}(X), E_{Q_2}(X))$, whereas $\delta_{\max}$ is the diameter of $\mathcal{X}$. □

4 Asymptotic knowledge acquisition and consensus formation as the number of data points increases

We will find conditions under which the posterior distribution P_i in (30), of each agent A_i, converges to a point mass $\delta_{x_{i\infty}}$ at $x_{i\infty} \in \mathcal{X}$ as $N \to \infty$. This limit $x_{i\infty}$ represents the ultimate beliefs of A_i for large datasets. Although we assumed in Section 3.1 that agents form their beliefs based on the whole data set $\boldsymbol{D}$ of size N, it is in fact possible to update the agents' posterior beliefs sequentially, if data points D_k arrive one by one for $k = 1, 2, \ldots$. For instance, if the posterior distributions P_i of all Agents A_i have been formed based on a data set of size N, and a new data point D_{N+1} arrives,

they could update their posteriors by communicating their likelihoods of the last data point D_{N+1} only. If the posterior belief of A_i is updated in this way as

$$P_i(x) \mapsto \frac{P_i(x)\prod_{j=1}^{n}[u_j(D_{N+1};x)f(D_{N+1};x)]^{w_{ji}M_{j,N+1}}}{Z_i} \tag{46}$$

when D_{N+1} arrives, for some normalizing contant Z_i, it follows from (30)-(31) that this is equivalent to A_i forming his beliefs for the whole augmented data set $D_1, \ldots, D_{N+1}$ according to (30). For this reason, it is meaningful to take a stochastic process point of view and speak of belief formation, knowledge acquisition and consensus formation sequentially as the size N of the data set increases.

To find the limit of all posterior distributions P_i as $N \to \infty$, it is convenient to introduce random variables $X_{iN} \sim P_i$ for $i = 1, \ldots, n$. These random variables need to be defined on the same probability space, so that the random vector $\boldsymbol{X}_N = (X_{1N}, \ldots, X_{nN})$ is well defined. This is achieved by assuming that

$$\{X_{iN}\}_{i=1}^{n} | \boldsymbol{D} \text{ are conditionally independent.} \tag{47}$$

The rationale for (47) is that $\boldsymbol{X}_N$ represents the random guesses of the value of x_0 of all agents, distributed in accordance with their respective posterior distributions $P_1, \ldots, P_n$. Recall from (35)-(36) that these posteriors are different functions $P_i(x) = P(x; \bar{\alpha}_i(\boldsymbol{D}))$ of the same data set $\boldsymbol{D}$, and therefore they are dependent. However, *conditionally* on $\boldsymbol{D}$, the n agents' random guesses of x_0 should be independent, in accordance with (47).

In order to deal with asymptotic KA and CF, we will generalize (10) from individuals ($n = 1$) to populations ($n > 1$). More specifically, we will prove weak convergence

$$\sqrt{N}(\boldsymbol{X}_N - \boldsymbol{x}_\infty) \xrightarrow{\mathcal{L}} N(0, V) \tag{48}$$

as $N \to \infty$, where

$$\boldsymbol{x}_\infty = (x_{1\infty}, \ldots, x_{n\infty}) \tag{49}$$

contains the limiting parameter values $x_{i\infty}$ of all agents A_i, whereas V is an asymptotic covariance matrix of order nq. If follows from (37) and (48) that the posterior distribution P of the whole population converges in probability to a mixture of point masses, since

$$P \xrightarrow{p} \frac{1}{n}\sum_{i=1}^{n} \delta_{x_{i\infty}}$$

as $N \to \infty$. It may happen that $x_{i\infty} \neq x_0$ for some agents A_i (and hence asymptotic full KA fails, since $\boldsymbol{x}_\infty \neq \boldsymbol{x}_0 = (x_0, \ldots, x_0)$) because of biased missing data mechanism for those agents that are influential for their beliefs. Let also C_N refer to the degree of consensus (42) within the population when the size of the data set is N. Under appropriate conditions we expect

$$C_N \xrightarrow{p} c_\infty \tag{50}$$

as $N \to \infty$. The less isolated the nodes of $\mathcal{A}$ are, or the more similar their data sets $\boldsymbol{D}_i$ are, the larger amount of limiting consensus c_∞ we expect.
To prove (48), we will assume that the missingness mechanism is not corrected for, according to (25), so that the posterior distribution $P_i(x)$ of agent A_i is given by (35). Introduce the prior score function

$$\begin{aligned} \psi_{i0}(x) &= P'_{i0}(x)/P_{i0}(x) \\ &= \textstyle\sum_{r=1}^{q} \eta'_r(x)\alpha_{ir} - \sum_{r=1}^{s} \alpha_{i,q+r} B'_r(x) \end{aligned} \tag{51}$$

of agent A_i and let $\boldsymbol{d} = (d_1, \ldots, d_N)$ be the complete data vector of size N. Recall also from (12) that $\psi(d_k; x)$ is the score function of each data point d_k. It follows from (29), (32), and (35) that the posterior score function of A_i can be expressed as a sum

$$\begin{aligned} \psi_i(x) &= \psi_i(\boldsymbol{d}; x) \\ &= d \log[P_i(x)]/dx \\ &= P'_i(x)/P_i(x) \\ &= \textstyle\sum_{j=1}^{n} w_{ji}\psi_{j0}(x) + \sum_{k=1}^{N}\sum_{j=1}^{n} w_{ji} M_{jk}\psi(d_k; x) \\ &= \bar{\psi}_{i0}(x) + \textstyle\sum_{k=1}^{N} \bar{M}_{ik}\psi(d_k; x) \end{aligned} \tag{52}$$

of two terms; a prior score function $\bar{\psi}_{i0}(x)$ and likelihood score function $\sum_{k=1}^{N} \bar{M}_{ik}\psi(d_k; x)$, both with contributions from all agents A_j that influence A_i. When N is large, only the latter likelihood-based score function will have an asymptotic impact. Because of the frequentist assumption made in Section 2.1, the components of the random data vector $\boldsymbol{D}$ are independent and identically distributed with density $f(\cdot; x_0)$. Since it is implicit that x_0 is the true parameter value, we let $E = E_{x_0}$ denote expectation of functions of $\boldsymbol{D}$. With this notation we deduce from (52) that the expected value of the posterior score function of A_i equals

$$\begin{aligned} E[\psi_i(\boldsymbol{D}; x)] &= \bar{\psi}_{i0}(x) + N\textstyle\sum_{j=1}^{n} w_{ji}\mu_j(x) \\ &=: \bar{\psi}_{i0}(x) + N\bar{\mu}_i(x), \end{aligned} \tag{53}$$

where

$$\mu_j(x) = E[M_j\psi(D; x)] = E[E(M_j|D)\psi(D; x)] = E[v_j(D)\psi(D; x)], \tag{54}$$

M_j is a binary variable that indicates whether agent A_j receives a single data point $D \sim f(\cdot; x_0)$ or not, and in the last step of (54) we made use of the MAR assumption (20). Equation (54) simplifies to the expected value $\mu_j(x) = \mu(x) = E[\psi(D; x)]$ of the score function (12) when A_j has no missing data, i.e., $P(M_j = 1) = 1$. We note from (53) that $N\mu_i(x)$ and $N\bar{\mu}_i(x)$ are the expected values of the likelihood part of A_i's posterior score function, after the first and second steps of belief formation.
Let $\mathrm{Cov} = \mathrm{Cov}_{x_0}$ refer to covariance between two functions of the random data vector $\boldsymbol{D}$ when x_0 is the true parameter. Since the missingness variables of different data points are independent (cf. (21)), it follows from the fourth line of (52) that the

covariance between the posterior scores of A_i and A_j, at parameters x_i and x_j, equals

$$\begin{aligned} \mathrm{Cov}[\psi_i(\boldsymbol{D};x_i),\psi_j(\boldsymbol{D};x_j)] &= N\sum_{l,m=1}^{n} w_{li}w_{mj}J_{lm}(x_i,x_j) \\ &=: N\bar{J}_{ij}(x_i,x_j) \end{aligned} \tag{55}$$

where

$$\begin{aligned} J_{lm}(x_i,x_j) &= \mathrm{Cov}[M_l\psi(D;x_i),M_m\psi(D;x_j)] \\ &= E[v_l(D)v_m(D)^{1(m\neq l)}\psi(D;x_i)^T\psi(D;x_j)] - \mu_l(x_i)\mu_m(x_j), \end{aligned} \tag{56}$$

making use of conditional independence (22) between M_l and M_m given D, in the last step. When A_l and A_m have no missing data, i.e. $P(M_l=1)=P(M_m=1)=1$, the matrix $J_{lm}(x,x)$ in (56) equals the Fisher information matrix $J(x)$ in (11). It follows from (55) that $NJ_{ij}(x_i,x_j)$ and $N\bar{J}_{ij}(x_i,x_j)$ correspond to the covariance between the likelihood-based score functions of agents A_i and A_j, after the first and second steps of belief formation.

Let $H_i(x)=H_i(\boldsymbol{d};x)=\psi_i'(\boldsymbol{d};x)$ be the Hessian (second order derivative) matrix of $\log[P_i(x)]$. By differentiating (53) with respect to x we find that

$$\begin{aligned} E[H_i(\boldsymbol{D};x)] &= \bar{\psi}_{i0}'(x) + N\sum_{j=1}^{n} w_{ji}\mu_j'(x) \\ &= \bar{\psi}_{i0}'(x) + N\bar{\mu}_i'(x) \end{aligned} \tag{57}$$

where

$$\mu_j'(x) = E[v_j(D)\psi'(D;x)].$$

With these preliminaries, we are ready to formulate the following result, which is proved in Appendix A:

Theorem 1 (Asymptotic beliefs.) *Suppose the n agents of a population receive N data items $\boldsymbol{D}=(D_1,\ldots,D_N)$ with a set of fixed mutual influences $\boldsymbol{W}=(w_{ji})$, not depending on data. Assume also that the agents do not correct their likelihoods for missing data, according to (25), so that the posterior distributions $P_i(x)$, of the q-dimensional parameter $x\in\mathcal{X}$, are given by (35) for $i=1,\ldots,n$. Let also $\boldsymbol{X}_N=(X_{1N},\ldots,X_{nN})$ be a collection of n random variables, whose components $X_{iN}\sim P_i$ are distributed as the posterior distributions and additionally satisfy (47). Then $\boldsymbol{X}_N$ is asymptotically normally distributed (48) as $N\to\infty$, with a limiting mean vector $\boldsymbol{x}_\infty$ in (49), whose components $x_{i\infty}$ solve the q-dimensional system of equations*

$$\bar{\mu}_i(x_{i\infty}) = 0 \tag{58}$$

for $i=1,\ldots,n$, with $\bar{\mu}_i(x)$ as defined in (53). Moreover, the asymptotic covariance matrix of (48) is given by

$$V = \bar{\mu}'(\boldsymbol{x}_\infty)^{-1}\bar{J}(\boldsymbol{x}_\infty)\bar{\mu}'(\boldsymbol{x}_\infty)^{-1} - \bar{\mu}'(\boldsymbol{x}_\infty)^{-1}, \tag{59}$$

where

$$\bar{J}(\boldsymbol{x}) = (\bar{J}_{ij}(x_i,x_j))_{i,j=1}^{n} \tag{60}$$

and

$$\bar{\mu}'(\boldsymbol{x}) = Diag(\bar{\mu}_1'(x_1),\ldots,\bar{\mu}_n'(x_n)) \tag{61}$$

are symmetric, square matrices of order nq, defined for any vector $\boldsymbol{x}=(x_1,\ldots,x_n)$ in $\mathcal{X}^n$, whereas $\bar{J}_{ij}(x_i,x_j)$ and $\bar{\mu}_i'(x)$ are defined in (55) and (57) respectively.

Theorem 1 has the following consequences in terms of asymptotic KA and CF:

Corollary 1 (Asymptotic KA.) *Suppose a population satisfies the conditions of Theorem 1. Agent A_i then acquires full knowledge about x_0 asymptotically at rate $\varepsilon_N = 1/\sqrt{N}$ if and only if $x_{i\infty} = x_0$. The whole population acquires full knowledge about x_0 asymptotically at rate $\varepsilon_N = 1/\sqrt{N}$ if and only if*

$$\boldsymbol{x}_\infty = (x_0, x_0, \ldots, x_0). \tag{62}$$

Corollary 2 (Asymptotic CF.) *Suppose the distance measure (43) is used for quantifying consensus. The population described in Theorem 1 then reaches full consensus asymptotically as $N \to \infty$, that is, (50) holds with*

$$c_\infty = 1, \tag{63}$$

if and only if

$$\boldsymbol{x}_\infty = (x, x, \ldots, x) \tag{64}$$

for some $x \in \mathcal{X}$.

We see from Corollaries 1-2 that asymptotic KA of a population is a stronger concept than asymptotic CF. Indeed, both concepts require that all elements of $\boldsymbol{x}_\infty$ are identical, whereas asymptotic KA additionally requires that these elements equal x_0. In order to find a more explicit condition for the asymptotic value $\delta_{x_{i\infty}}$ of Agent A_i's posterior distribution in Theorem 1, we will rephrase (58) for exponential family data. This is dealt with in the following corollary:

Corollary 3 (Explicit expression for $x_{i\infty}$.) *Suppose data of Theorem 1 are obtained from an exponential family (4), with a vector*

$$T(d) = (T_1(d), \ldots, T_q(d)) \tag{65}$$

of sufficient statistics. The q-dimensional system of equations (58) for finding the asymptotic value $x_{i\infty}$ of agent A_i's posterior beliefs can then be rephrased as

$$E_{x_{i\infty}}[T(D)] = \sum_{j=1}^{n} \tilde{w}_{ji} \tilde{E}_j[T(D)] \tag{66}$$

where

$$\tilde{w}_{ji} = \frac{w_{ji} v_j}{\sum_{r=1}^{n} w_{ri} v_r} \tag{67}$$

are effective influence weights that not only take the original influence weights w_{ji} into account, but also the probabilities

$$v_j = E[v_j(D)] \tag{68}$$

of A_j not missing a data point $D \sim f(\cdot; x_0)$. Moreover,

$$\tilde{E}_j[T(D)] = \frac{E[T(D) v_j(D)]}{v_j} = \int_{\mathcal{D}} T(z) \tilde{f}_j(z; x_0) dz \tag{69}$$

is a size-biased estimate of $E[T(D)]$, where $D \sim \tilde{f}_j(\cdot; x_0)$ has density function

$$\tilde{f}_j(d; x_0) = \frac{v_j(d) f(d; x_0)}{\int_{\mathcal{D}} v_j(z) f(z; x_0) dz}. \tag{70}$$

Remark 2 (Explicit expression for $\boldsymbol{x}_\infty$.) Equation (66) is an explicit formula for obtaining the asymptotic limit $\delta_{x_{i\infty}}$ of agents A_i's posterior distribution P_i. In order to formulate this condition jointly for all agents $A_1, \ldots, A_n$, it is convenient to introduce the q-dimensional function

$$\rho(x) = E_x[T(D)] \tag{71}$$

for the expected sufficient statistic vector at all possible parameter values $x \in \mathcal{X}$, as well as the nq-dimensional function

$$\boldsymbol{\rho}(\boldsymbol{x}) = (\rho(x_1), \ldots, \rho(x_n)) \tag{72}$$

for all $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$. With this notation we can reexpress (66) jointly for all agents. This is formulated as an nq-dimenional system of equations

$$\boldsymbol{\rho}(\boldsymbol{x}_\infty) = \tilde{\boldsymbol{\rho}}\tilde{\boldsymbol{W}} \tag{73}$$

for finding $\boldsymbol{x}_\infty$ in (49), where

$$\tilde{\boldsymbol{W}} = (\tilde{w}_{ji})_{j,i=1}^n = \mathrm{Diag}(\boldsymbol{v})\boldsymbol{W}\mathrm{Diag}(\boldsymbol{v}\boldsymbol{W})^{-1} \tag{74}$$

is the collection of all effective influence weights (67), $\boldsymbol{W}$ is the corresponding matrix of influence weights w_{ji} in (13), $\boldsymbol{v} = (v_1, \ldots, \boldsymbol{v}_n)$ contains all probabilities (68) of retaining a data point for all agents, whereas

$$\begin{aligned} \tilde{\boldsymbol{\rho}} &= (\tilde{\rho}_1, \ldots, \tilde{\rho}_n) \\ &= (\tilde{E}_1[T(D)], \ldots, \tilde{E}_n[T(D)]) \end{aligned} \tag{75}$$

contains the size-biased expected sufficient statistics (69) of all agents. □

Remark 3 (Modified influence weights.) We see from (66)–(67) that an agent A_j who discards data (lowers his v_j) will also lower his effective influence $\tilde{w}_{ji}$ on agent A_i. It is possible though for A_j to compensate for this by replacing his first step posterior distribution in (28) as

$$P_{j1}(x) \mapsto \frac{P_{j1}(x)^{t_j}}{Z_j}, \tag{76}$$

for some normalizing constant Z_j. To see this, notice that the parameter $t_j > 0$ will determine how certain A_j is about his beliefs, that is, how concentrated $P_{j1}(x)$ is around its mode $\hat{x}_{j1}$. In particular, $t_j > 1$ ($0 < t_j < 1$) corresponds to agent A_j being more (less) certain about his beliefs in $\hat{x}_{j1}$ than a traditional Bayesian inference would allow. If, for instance, data is normally distributed according to Example 3 and $P_{j1} \sim N(\hat{x}_{j1}, \sigma_{j1}^2)$, then (76) modifies A_j's first step posterior belief to $P_{j1} \sim N(\hat{x}_{j1}, \sigma_{j1}^2/t_j)$. If on the other hand, data follows a Bernoulli distribution (see Example 1) and Agent A_j has a first step beta posterior $P_{ji} \sim B(a_{j1}, b_{j1})$, then (76) modifies A_j's first step posterior belief to another beta distribution, $P_{ji} \sim B(t_j a_{j1} - t_j + 1, t_j b_{j1} - t_j + 1)$, with the same mode.

It can be seen from (30) that the way in which A_i combines the first step posteriors $P_{11}, \ldots, P_{n1}$ of all agents, in order to update his own posterior from P_{i1} to $P_i = P_{i2}$, is such that (76) changes A_j's influence on A_i from w_{ji} to $t_j w_{ji}$, whereas A_j's effective influence on A_i gets proportional to $v_j t_j w_{ji}$ instead of $v_j w_{ji}$. In particular, a choice $t_j = v_j^{-1}$ in (76) would compensate A_j for loosing influence because of missing data, so that his effective influence on A_i is proportional to w_{ji}. However, if missingness mechanism (24) is used with $s_j > 0$, then typically v_j will depend on the unknown x_0, and hence v_j is unknown to A_j. Agent A_j can still use $t_j = 1/\hat{v}_j$, where $\hat{v}_j = |\boldsymbol{D}_j|/N$ is an asymptotically consistent estimate of v_j. □

Remark 4 (Properties of V.) It is of interest to analyze how the weight matrix $\boldsymbol{W} = (w_{ji})$ affects the asymptotic covariance matrix (59) of Theorem 1. Let

$$V = (V_{ij})_{i,j=1}^{n} \tag{77}$$

be a decomposition of the covariance matrix in (59) into blocks, where V_{ij} is a square matrix of order q that corresponds to the asymptotic covariance between the beliefs of agents A_i and A_j. Also write $V = V_1 + V_2$, where $V_1 = (V_{1ij})_{i,j=1}^{n}$ and $V_2 = (V_{2ij})_{i,j=1}^{n}$ represent the two terms of (59). Suppose the weighs of influence for A_i are renormalized according to (16) for some constant a_i. It is shown in Appendix A that V_1 is unaffected by the modified weights, whereas the i-th block diagonal element of the block diagonal matrix V_2 changes from V_{2ii} to V_{2ii}/a_i. This is well in line with Remark 1, since V_1 is the asymptotic covariance matrix of the maximum aposteriori (MAP) estimator $\hat{\boldsymbol{x}} = (\hat{x}_1, \ldots, \hat{x}_n)$ of x_0 for all agents, whereas V_2 quantifies aposteriori uncertainty around $\hat{\boldsymbol{x}}$ for all agents. Since the MAP estimator is unaffected by renormalized weights (16), this renormalization will not change V_1.

We say that agents A_i and A_j, with $j \neq i$, are unconneted if they have no common influencer A_l. This means that

$$w_{li} w_{lj} = 0 \text{ for } l = 1, \ldots, n. \tag{78}$$

It follows from (55) and (59) that typically $\bar{J}_{ij}(x_{i\infty}, x_{j\infty}) \neq 0$ and $V_{ij} \neq 0$, also for a pair (A_i, A_j) of unconnected agents. That is, their random guesses of the value of x_0 are typically dependent, in spite of being unconnected, since they still make use of the same data $\boldsymbol{D} = (D_1, \ldots, D_N)$. If, on the other hand, all agents' data points D_{ik}, at each time point k, would be different and independent for $i = 1, \ldots, n$, then unconnected agents' guesses of x_0 would be independent as well. □

As a next step, we apply Theorem 1 to the MCAR missingness mechanism described in Example 9:

Corollary 4 (Data missing completely at random.) *Suppose data is missing completely at random in Theorem 1, with a missingness probability v_i in (23) for agent A_i, $i = 1, \ldots, n$. The asymptotic normality result (48) then holds with a mean vector (49) whose components satisfy*

$$x_{i\infty} = x_0, \quad i = 1, \ldots, n \tag{79}$$

whereas the covariance matrix in (59) and (77) has blocks

$$V_{ij} = \frac{\bar{v}_i \bar{v}_j + \bar{v}_i 1(i=j) + \sum_{l=1}^{n} w_{li} w_{lj} v_l (1 - v_l)}{\bar{v}_i \bar{v}_j} J^{-1}(x_0) \tag{80}$$

of order q, with

$$\bar{v}_i = \sum_{j=1}^{n} w_{ji} v_j \tag{81}$$

and $J(x)$ the Fisher information matrix in (11). In particular, the population acquires knowledge asymptotically about x_0 at rate $1/\sqrt{N}$, and reaches full consensus asymptotically.

Remark 5 For single individuals ($n = 1$) with MCAR data we have $\bar{v}_i = v_i = v$ in Corollary 4, where v is the probability that the single agent does not miss data. The asymptotic covariance matrix in (59) then simplifies to

$$V = 2v^{-1} J^{-1}(x_0). \tag{82}$$

Note that V in (82) agrees with the asymptotic covariance matrix of (10) when no data is missing, i.e. $v = 1$. □

5 Iterated updates of beliefs

In Section 3.1, we assumed that the posterior distribution of each agent A_i in the population is updated from the prior distribution P_{i0} in (29), in two steps. In Step 1, A_i forms a posterior distribution P_{i1} based on his own data (cf. (28)), whereas in Step 2 he updates his posterior from P_{i1} to P_{i2} based on inputs from the other agents, according to (30). In this section, we will assume that the posterior distributions of all agents are updated L times, based on inputs from other agents, as

$$P_{il}(x) \propto \prod_{j=1}^{n} P_{j,l-1}(x)^{w_{ji}} \text{ for } l = 2, \ldots, L+1. \tag{83}$$

This is a natural extension of the DeGroot model [8], in that posterior distributions (rather than random guesses of elements in $\mathcal{X}$) are iterated.
The following proposition implies that L rounds of iteration of the posterior distribution, among the agents of the population, is equivalent to using a matrix

$$\boldsymbol{W}^L = (w_{ji}^{(L)})_{j,i=1}^{n} \tag{84}$$

of influence weights, with $\boldsymbol{W}$ as in (13).

Proposition 1 (Iterated updates of posterior distribution.) *Suppose all agents A_i of the population update their first step posteriors P_{i1} in (28), by collecting information about the other agents' posterior distributions, according to (83), a total of L times. The posterior distribution of A_i is then given by*

$$\begin{aligned} P_{i,L+1}(x) &= \textstyle\prod_{j=1}^{n} P_{j1}(x)^{w_{ji}^{(L)}} \\ &\propto \textstyle\prod_{j=1}^{n} P_{j0}(x)^{w_{ji}^{(L)}} \cdot \prod_{k=1}^{N} \prod_{j=1}^{n} [u_j(D_k;x) f(D_k;x)]^{w_{ji}^{(L)} M_{jk}} \\ &= \textstyle\prod_{j=1}^{n} P_{j0}(x)^{w_{ji}^{(L)}} \prod_{k=1}^{N} [\bar{u}_{ik}^{(L)}(D_k;x) f(D_k;x)]^{\bar{M}_{ik}^{(L)}}, \end{aligned} \tag{85}$$

with influence weights $w_{ji}^{(L)}$ as in (84),

$$\bar{M}_{ik}^{(L)} = \sum_{j=1}^{n} w_{ji}^{(L)} M_{jk},$$

and

$$\bar{u}_{ik}^{(L)}(D_k;x) = \left[\prod_{j=1}^{n} u_j(D_k;x)^{w_{ji}^{(L)} M_{jk}} \right]^{1/\bar{M}_{ik}^{(L)}}.$$

5.1 Large N asymptotics

It is possible (as in Section 4 for $L = 1$) to formulate belief formation after L rounds of iterations sequentially when data points D_k arrive one by one for $k = 1, 2, \ldots$. Suppose Agent A_i has a posterior distribution $P_{i,L+1}$ after L rounds of iteration, based on a

data set of size N. If a new data point D_{N+1} arrives, it follows from (85) that it is possible for A_i to update his posterior distribution as

$$P_{i,L+1}(x) \mapsto \frac{P_{i,L+1}(x)\prod_{j=1}^{n}[u_j(D_{N+1};x)f(D_{N+1};x)]^{w_{ji}^{(L)}M_{j,N+1}}}{Z_i} \tag{86}$$

for some normalizing contant Z_i. Equation (86) corresponds to a situation where agents update their posteriors when D_{N+1} arrives by communicating among each other, and updating, their likelihoods of D_{N+1} a total of L times. Therefore, it is meaningful to speak of belief formation for a fixed L, sequentially as the size N of the data set increases. With this in mind, we obtain the following consequence of Theorem 1, Proposition 1, Remark 2, and Corollaries 1–2:

Theorem 2 (Asymptotic beliefs for iterated inputs of beliefs.) *Suppose the conditions of Theorem 1 hold, with the difference that the n agents of the population iterate their original posterior beliefs $P_{11},\dots,P_{n1}$ a total of L times, according to (83). Let $\boldsymbol{X}_N^{(L)} = (X_{1N}^{(L)},\dots,X_{nN}^{(L)})$ be a collection of n random variables, whose components $X_{iN}^{(L)} \sim P_{i,L+1}$ are distributed as the posterior distributions of all agents for a data set of size N after L steps of belief iteration. Suppose also that the components of $\boldsymbol{X}_N^{(L)}$ satisfy (47) (with $X_{iN}^{(L)}$ in place of X_{iN}). Then $\boldsymbol{X}_N^{(L)}$ is asymptotically normally distributed*

$$\sqrt{N}(\boldsymbol{X}_N^{(L)} - \boldsymbol{x}_\infty^{(L)}) \xrightarrow{\mathcal{L}} N(0, V^{(L)}) \tag{87}$$

as $N\to\infty$, where $\boldsymbol{x}_\infty^{(L)} = (x_1^{(L)},\dots,x_n^{(L)})$ and $V^{(L)}$ are defined analogously to $\boldsymbol{x}_\infty$ and V in (58) and (59) respectively, with influence weights $w_{ji}^{(L)}$ in (84) instead of w_{ji} in (13). If additionally data is drawn from an exponential family (4) with q natural parameters, the vector $\boldsymbol{x}_\infty^{(L)}$ in (87) solves the nq-dimensional system

$$\boldsymbol{\rho}(\boldsymbol{x}_\infty^{(L)}) = \tilde{\boldsymbol{\rho}}\tilde{\boldsymbol{W}}^{(L)} \tag{88}$$

of equations, with $\boldsymbol{\rho}(\boldsymbol{x})$ defined in (71)–(72),

$$\tilde{\boldsymbol{W}}^{(L)} = (\tilde{w}_{ji}^{(L)})_{j,i=1}^{n} = Diag(\boldsymbol{v})\boldsymbol{W}^L Diag(\boldsymbol{v}\boldsymbol{W}^L)^{-1} \tag{89}$$

the collection of effective influence weights after L rounds of belief iterations, whereas $\tilde{\boldsymbol{\rho}}$ and $\boldsymbol{v}$ are defined in and above (75) respectively. The population acquires knowledge about x_0 asymptotically at rate $\varepsilon_N = 1/\sqrt{N}$ if and only if

$$\boldsymbol{x}_\infty^{(L)} = (x_0, x_0, \dots, x_0). \tag{90}$$

If a distance measure (43) between posterior distributions is used to quantify consensus, the population reaches full consensus asymptotically ($C = C_N^{(L)} \to c_\infty^{(L)} = 1$ in (42) as $N\to\infty$), if

$$\boldsymbol{x}_\infty^{(L)} = (x, x, \dots, x) \tag{91}$$

for some $x \in \mathcal{X}$.

5.2 Large L asymptotics

We will next investigate what happens when $L \to \infty$, under the condition (15) that all weights of influence on A_i from other agents sum to 1. With this assumption the transpose $\boldsymbol{W}^T$ of the weight matrix (13) is the transition matrix of a Markov chain. We will therefore make use the asymptotic theory of Markov chains to study the $L \to \infty$ limit of KA and CF. This can be viewed as an extension of the DeGroot's consensus formation model [8], where agents iteratively update their beliefs as linear combinations of other agents' beliefs, according to a transition matrix $\boldsymbol{W}^T$ of a Markov chain. In our context $\boldsymbol{W}^T$ is rather used to model how each agent weights the posterior distributions from other agents. In order to formulate an asymptotic KA and CF result for $L \to \infty$, we first need some concepts.

Definition 5 (Classification of agents.) A collection $\mathcal{B} \subset \mathcal{A}$ of agents is **closed** if $w_{ji}^{(l)} = 0$ whenever $j \notin \mathcal{B}$ and $i \in \mathcal{B}$, for all $l = 1, 2, \ldots$. Two agents A_i and A_j have a **two-sided communication** if there exist $l_1 \geq 1$ and $l_2 \geq 1$ such that $w_{ij}^{(l_1)} > 0$ and $w_{ji}^{(l_2)} > 0$ respectively. A set $\mathcal{B}$ of agents is **irreducible** if all agents in $\mathcal{B}$ communicate two-sidedly. An agent A_i is **recurrent** or **transient** depending on whether $\sum_{l=1}^{\infty} w_{ii}^{(l)}$ is infinite or finite, respectively. The **period** $p(i)$ of agent A_i is the greatest common divisor of all l such that $w_{ii}^{(l)} > 0$. In particular, A_i is **aperiodic** if $p(i) = 1$. □

With these definitions we can formulate the following result:

Theorem 3 (Asymptotic weight matrix and consensus when $L \to \infty$.) *Suppose weight matrix $\boldsymbol{W} = (w_{ji})$ of influences in (13) satisfies condition (15). It is possible then to divide the population $\mathcal{A}$ into $m + 1$ disjoint components, as*

$$\mathcal{A} = \mathcal{T} \cup \mathcal{B}_1 \cup \ldots \cup \mathcal{B}_m, \tag{92}$$

where $\mathcal{T}$ is the set of transient agents, whereas $\mathcal{B}_t$ is closed and irreducible for $t = 1, \ldots, m$. All agents of $\mathcal{B}_t$ have the same period $p(\mathcal{B}_t)$. If the closed and irredicuble components of $\mathcal{A}$ are aperiodic, i.e.

$$p(\mathcal{B}_1) = \ldots = p(\mathcal{B}_m) = 1,$$

then

$$\boldsymbol{W}^L \to \boldsymbol{W}^{(\infty)} = (w_{ji}^{(\infty)})_{j,i=1}^{n} = ((\boldsymbol{w}_1^{(\infty)})^T, \ldots, (\boldsymbol{w}_n^{(\infty)})^T) \text{ as } L \to \infty, \tag{93}$$

where $\boldsymbol{w}_i^{(\infty)} = (w_{1i}^{(\infty)}, \ldots, w_{ni}^{(\infty)})$ contains the asymptotic influences of all agents on A_i, with $w_{ji}^{(\infty)} \geq 0$ and $\sum_{j=1}^{n} w_{ji}^{(\infty)} = 1$. Moreover,

$$\boldsymbol{w}_i^{(\infty)} = \boldsymbol{b}_t \text{ for all } A_i \in \mathcal{B}_t, \tag{94}$$

where $\boldsymbol{b}_t = (b_{t1}, \ldots, b_{tn})$ has $b_{tj} > 0$ and $b_{tj} = 0$ for all j such that $A_j \in \mathcal{B}_t$ and $A_j \notin \mathcal{B}_t$ respectively, and $\sum_{j \in \mathcal{B}_t} b_{tj} = 1$. In addition,

$$\boldsymbol{w}_i^{(\infty)} = \sum_{t=1}^{m} \alpha_{it} \boldsymbol{b}_t \text{ for each } A_i \in \mathcal{T}, \tag{95}$$

where $\alpha_{it} \geq 0$ and $\sum_{t=1}^{m} \alpha_{it} = 1$. In particular, consensus is always obtained asymptotically as $L \to \infty$ when $m = 1$, in the sense that the consensus variable C in (42) satisfies

$$C = C^{(L)} \to c^{(\infty)} = 1. \tag{96}$$

We refer to $\boldsymbol{b}_t$ as the long-run influence vector of component $\mathcal{B}_t$, with b_{tj} quantifying the long-run influence that agent A_j has on belief formation in $\mathcal{B}_t$. Intuitively, (94) tells that agents within each irreducible component $\mathcal{B}_t$ are only influenced by each other as $L \to \infty$, whereas (95) implies that all transient agents lose their influence asymptotically as L grows. This is so, since a transient agent A_j has $b_{tj} = 0$ for $t = 1, \ldots, m$. In order to illustrate Theorem 3, we will now return to the population structures that were introduced in Section 3.1.1.

Example 13 (Symmetric population, contd.) Let us start by renormalizing the influence weights of Example 5 so that (15) holds, i.e.

$$w_{ji} = \frac{1(i = j) + w1(i \neq j)/(n-1)}{1 + w}, \tag{97}$$

for some $w \geq 0$. When $w > 0$ in (97), all agents communicate two-sidedly, and we will find below that consensus is attained asymptotically as $L \to \infty$. It turns out that the value of $w > 0$ does not affect the limiting consensus element of $\mathcal{X}$, only the rate at which consensus is attained. In order to verify this, we make use of Theorem 3 and notice that the whole population forms one single, irreducible component ($\mathcal{A} = \mathcal{B}$) when $w > 0$ in (97), with a uniform asymptotic distribution $\boldsymbol{b} = (1/n, \ldots, 1/n)$. Hence $\boldsymbol{W}^{(\infty)} = (w_{ij}^{(\infty)})_{j,i=1}^{n}$ has identical elements $w_{ji}^{(\infty)} = 1/n$. This in turn implies that the effective influence matrix $\tilde{\boldsymbol{W}}^{(\infty)} = (\tilde{w}_{ji}^{(\infty)})_{j,i=1}^{n}$ in (89) has identical columns, with

$$\tilde{w}_{ji}^{(\infty)} = \frac{v_j}{\sum_{r=1}^{n} v_r}. \tag{98}$$

In view of (88) and (91), it follows that asymptotic consensus $\boldsymbol{x}_{\infty}^{(\infty)} = (x, \ldots, x)$ is attained, where x solves the equation

$$\rho(x) = \frac{\sum_{r=1}^{n} \tilde{\rho}_r v_r}{\sum_{r=1}^{n} v_r}. \tag{99}$$

Full asymptotic knowledge is only attained if $x = x_0$ in (99). This happens, for instance, for data MCAR, with $\tilde{\rho}_i = \rho(x_0)$ for $i = 1, \ldots, n$.
When $w = 0$, all agents are isolated. This corresponds to a decomposition of the population into n isolated irreducible components, $\mathcal{B}_i = \{A_i\}$ for $i = 1, \ldots, n$. This implies that $\boldsymbol{W}^{(\infty)} = \tilde{\boldsymbol{W}}^{(\infty)}$ are both identity matrices of order n. The components of the asymptotic belief vector $\boldsymbol{x}_{\infty}^{(\infty)} = (x_1, \ldots, x_n)$ then solve the equations

$$\rho(x_i) = \tilde{\rho}_i$$

for $i = 1, \ldots, n$. Asymptotic consensus is typically *not* attained, unless $\tilde{\rho}_1 = \ldots = \tilde{\rho}_n$. This happens for data MCAR, or for biased missing data (24), where all agents A_i have the same preferred $y_i = y \in \mathcal{X}$, that they hold at the same strength $s_i = s$. Asymptotic KA additionally requires that $y = x_0$. □

Example 14 (Star-shaped population, contd.) For the star-shaped population of Example 6, we renormalize weights as

$$w_{ji} = \frac{1(i = j) + w1(j = 1, i \neq 1)}{1 + w1(i \neq 1)}, \tag{100}$$

so that (15) holds. Recall that this population has one influencer A_1, whose impact is stronger the larger $w > 0$ is in (100). We will find below that consensus is attained asymptotically as

$L \to \infty$. Similarly as in Example 13, $w > 0$ does not affect the limiting consensus parameter in $\mathcal{X}$, only that rate at which this consensus is attained. In order to demonstrate this we make use of Theorem 3 and notice that the population has one single recurrent agent $\mathcal{B} = \{A_1\}$, whereas all other agents are transient, i.e. $\mathcal{T} = \{A_2, \dots, A_n\}$. Consequently $\boldsymbol{W}^{(\infty)} = \tilde{\boldsymbol{W}}^{(\infty)}$ have elements $\tilde{w}_{ji}^{(\infty)} = 1(j = 1)$. It follows from (88) that consensus is attained asymptotically as $L \to \infty$, i.e. $\boldsymbol{x}_\infty^{(\infty)} = (x, \dots, x)$, where x solves the equation

$$\rho(x) = \tilde{\rho}_1. \tag{101}$$

From this, we deduce that the consesus value x of the population is determined solely by the influencer A_1. Knowledge is not attained asymptotically if the influencer A_1 has a biased missing data mechanism (24), with a preferred parameter value $y_1 \neq x_0$ at strength $s_1 > 0$. □

Example 15 (Linearly ordered population, contd.) Consider the linearly ordered subpopulation of Example 7, and renormalize weights as

$$w_{ji} = \frac{1(i = j) + w1(i = j + 1)}{1 + w1(i \neq 1)},$$

so that (15) holds. The asymptotic behaviour of the population, as the number L of iterations grows, is the same as in Example 14. That is, the population has one single recurrent agent $\mathcal{B} = \{A_1\}$, whereas all other agents are transient ($\mathcal{T} = \{A_2, \dots, A_n\}$). The asymptotic effective influence matrix $\tilde{\boldsymbol{W}}^{(\infty)}$ is the same as in Example 14, and consequently consensus $\boldsymbol{x}_\infty^{(\infty)} = (x, \dots, x)$ is attained as $L \to \infty$, where x solves Equation (101), and hence is determined only by the influencer A_1. □

Example 16 (Subdivided population, contd.) Recall that the subdivided population of Example 8 is divided into m subpopulations of equal size r, so that $n = mr$. We start by renormalzing influence weights as

$$w_{ji} = \frac{1(i = j) + w_1 1(i \neq j, [(i-1)/r] = [(j-1)/r]) + w_2 1([(i-1)/r] \neq [(j-1)/r])}{1 + (r-1)w_1 + r(m-1)w_2}, \tag{102}$$

so that (15) holds. Here $w_2 \geq 0$ and $w_1 > w_2$ quantify strength of influence between and within subpopulations, respectively. When $w_2 > 0$, all agents communicate two-sidedly. This implies that Theorem 3 holds with the whole population forming one single irreducible component ($\mathcal{A} = \mathcal{B}$), and with $\tilde{w}_{ji}^{(\infty)} = 1/n$, as in the first part of Example 13. It follows that asymptotic consensus $\boldsymbol{x}_\infty^{(\infty)} = (x, \dots, x)$ is attained as $L \to \infty$, where x solves (99).

When $w_2 = 0$ in (102), all subpopulations are isolated. When the number L of iterations grows, this implies that Theorem 3 is applicable with m irreducible components $\mathcal{B}_t = \{A_{(t-1)r+1}, \dots, A_{tr}\}$ for $t = 1, \dots, m$. By symmetry we see that the asymptotic influence distribution $\boldsymbol{b}_t = (b_{t1}, \dots, b_{tn})$ within component $\mathcal{B}_t$ is uniform, i.e. $b_{ti} = 1(i \in \mathcal{B}_t)/r$. Since all columns of $\boldsymbol{W}^{(\infty)}$ that correspond to $\mathcal{B}_t$ equal $\boldsymbol{b}_t^T$, we deduce from (88) that the asymptotic beliefs of all agents is represented by the vector

$$\boldsymbol{x}_\infty^{(\infty)} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_m),$$

where $\boldsymbol{x}_t = (x_t, \dots, x_t)$ is a vector of length r for $t = 1, \dots, m$, whose components solve the equation

$$\rho(x_t) = \frac{\sum_{i \in \mathcal{B}_t} \tilde{\rho}_i v_i}{\sum_{i \in \mathcal{B}_t} v_i}. \tag{103}$$

We see from (103) that when $L \to \infty$ and $w_2 = 0$ in (102), asymptotic consensus is obtained locally within subpopulations, but typically *not* over the whole population. □

5.3 Simultaneous asymptotics for L and N

In Examples 13–16 we derived the limiting belief vector $\boldsymbol{x}_\infty^{(\infty)}$ when the number of data points N as well as the number of iterations L among the agents are both large. It is possible to interpret $\boldsymbol{x}_\infty^{(\infty)}$ as follows: When data points D_k arrive sequentially, we assumed in (46) and (86) that agents sequentially communicate the likelihoods of their latest received data points, either 1 or L times. Another possibility is for the agents communicate their posterior distributions after every β:s data points, for some positive integer β. Since each posterior distribution involves the likelihoods of the data points received so far, this implies that the likelihoods of earlier received data points will be communicated more often that those that arrived later on. Indeed, suppose $N = L\beta + \gamma$, where L is another positive integer and $0 \le \gamma \le L - 1$. By generalizing the proof of Proposition 1, it can be seen that the posterior distribution $P_i = P_{iN}$ of Agent A_i satisfies

$$P_i(x) = \prod_{j=1}^{n} P_{j0}(x)^{w_{ji}^{(L)}} \cdot \prod_{k=1}^{N} \prod_{j=1}^{n} [u_j(D_k; x) f(D_k; x)]^{w_{ji}^{(L-[k/\beta])} M_{jk}} \tag{104}$$

where $[x]$ is the integer part of x and $w_{ji}^{(0)} = 1(i = j)$. It follows from (104) that the number of iterations $L = [N/\beta]$ of the priors as well as the number of iterations $L - [k/\beta] = [N/\beta] - [k/\beta]$ of each data point D_k diverge as $N \to \infty$. For exponential family models (4), the limiting vector beliefs of all agents, will therefore converge to an nq-dimensional vector $\boldsymbol{x}_\infty^{(\infty)}$ that solves the system of equations

$$\boldsymbol{\rho}(\boldsymbol{x}_\infty^{(\infty)}) = \tilde{\boldsymbol{\rho}} \tilde{\boldsymbol{W}}^{(\infty)}. \tag{105}$$

6 Examples of biased missing data

In this section, we will investigate how the biased missingness mechanism (24) affects the asymptotic limit $\boldsymbol{x}_\infty^{(L)}$ in (88), of all agents' posterior distributions. We will concentrate on Examples 1 and 3 of Section 2.1, where $x \in \mathcal{X}$ refers to a location parameter. Note that (24) is less appropriate to use for Examples 2 and 4, where $x \in \mathcal{X}$ involves scale parameters. For instance, for the exponential distribution of Example 2, if a value $y_j > 0$ of the scale parameter is preferred by agent A_j with strength $s_j > 0$, the missingness mechanism (24) will lead to a *smaller* fraction of missing data for all $x > y_j$.

6.1 Bernoulli distribution

We know from Example 1 that the Bernoulli distribution $D \sim \text{Be}(x)$ belongs to an exponential family (4) with $q = 1$ and $T(d) = d$. Formula (71) then simplifies to

$$\rho(x) = E_x[D] = x. \tag{106}$$

Recall that data missingness mechanism (24) means that agent A_j prefers parameter value $y_j \in \mathcal{X} = [0, 1]$ with strength s_j. Let

$$\theta_j = \left(\frac{y_j}{1 - y_j}\right)^{s_j}$$

be the odds ratio of the preferred parameter value of A_j, raised to a power of s_j. Since $\mathcal{D}$ is discrete, dz is the counting measure in the definition (70) of the size-biased data distribution $\tilde{f}_j$ of A_j, so that the integral of this equation corresponds to a sum. It follows from (24), after some computations, that this size-biased data distribution of A_j is

$$\tilde{f}_j(\cdot; x_0) \sim \text{Be}\left(\frac{\theta_j x_0}{1 - x_0 + \theta_j x_0}\right),$$

whereas the probability (68) that A_j retains a data point is

$$v_j = v_j^{max}\left[\min(\theta_j^{-1}, 1)(1 - x_0) + \min(\theta_j, 1)x_0\right]. \tag{107}$$

The size-biased estimate (75) for A_j, of the expected sufficient statistic $E[T(D)] = E(D) = x_0$, is

$$\tilde{\rho}_j = \frac{\theta_j x_0}{1 - x_0 + \theta_j x_0} = \frac{x_0}{\theta_j^{-1}(1 - x_0) + x_0} =: g(x_0, \theta_j), \tag{108}$$

which is less than, equal to, and larger than x_0 if $\theta_j < 1$, $\theta_j = 1$, and $\theta_j > 1$ respectively. Putting things together, it follows from (106) and (108) that (88) reduces to

$$\boldsymbol{x}_\infty^{(L)} = (g(x_0, \theta_1), \ldots, g(x_0, \theta_n))\tilde{\boldsymbol{W}}^{(L)}, \tag{109}$$

where $\tilde{\boldsymbol{W}}^{(L)}$ is the matrix (89) of effective influence weights $\tilde{w}_{ji}^{(L)}$. Note in particular that the asymptotic belief of agent A_i (component i of $\boldsymbol{x}_\infty^{(L)}$) is a weighted average

$$x_{\infty i}^{(L)} = \sum_{j=1}^{n} g(x_0, \theta_j)\tilde{w}_{ji}^{(L)}$$

of all $g(x_0, \theta_j)$ in (108). In order to obtain the vector $\boldsymbol{x}_\infty^{(\infty)}$ in (105), recall that values of the weights $\tilde{w}_{ji}^{(L)}$, as $L \to \infty$, are obtained for different population structures in Examples 13–16. For instance, $\tilde{w}_{ij}^{(\infty)}$ is given by (98) for the symmetric and connected

($w > 0$) population of Example 13, whereas the influencer A_1 receives all weight ($\tilde{w}_{1i}^{(\infty)} = 1$) for the star-shaped and linearly ordered populations of Examples 14–15.

6.2 Normal distribution with known variance

Recall from Example 3 that the normal distribution $D \sim N(x, \sigma^2)$, with σ^2 known, is an exponential family model (4) with $q = 1$ and $T(d) = d$. Consequently, (106) holds also for the normal distribution. If agent A_j prefers parameter value $y_j \in \mathcal{X} = \mathbb{R}$, with strength s_j, the size-biased distribution (70) of retained data points is

$$\tilde{f}_j(\cdot; x_0) \sim N\left(\frac{s_j y_j + x_0}{s_j + 1}, \frac{\sigma^2}{s_j + 1}\right).$$

The corresponding size-biased expected sufficient statistic (75) of agent A_j, is a weighted average

$$\tilde{\rho}_j = \frac{s_j}{s_j + 1} y_j + \frac{1}{s_j + 1} x_0 =: g(x_0; y_j, s_j), \tag{110}$$

of y_j and x_0, whereas the probability (68) that A_j retains a data point is

$$v_j = \frac{v_j^{\max}}{\sqrt{s_j + 1}} \exp\left[-\frac{s_j}{2\sigma^2(s_j + 1)}(y_j - x_0)^2\right]. \tag{111}$$

Putting things together, we deduce from (106) and (110) that (88) reduces to

$$\boldsymbol{x}_{\infty}^{(L)} = (g(x_0, y_1, s_1), \ldots, g(x_0, y_n, s_n))\tilde{\boldsymbol{W}}^{(L)}, \tag{112}$$

where $\tilde{\boldsymbol{W}}^{(L)}$ is the matrix (89) of effective influence weights $\tilde{w}_{ji}^{(L)}$. Note in particular that the asymptotic belief of agent A_i is a weighted average

$$x_{\infty i}^{(L)} = \sum_{j=1}^{n} g(x_0, y_j, s_j)\tilde{w}_{ji}^{(L)}$$

of all $g(x_0, y_j, s_j)$ in (110). As mentioned at the end of Section 6.1, explicit values of $\tilde{w}_{ji}^{(L)}$ are obtained for different population structures in Examples 13–16 as $L \to \infty$. This implies, for instance, that $x_{\infty i}^{(\infty)} = g(x_0, y_1, s_1)$ for the star-shaped and linearly ordered populations of Examples 14–15.

7 Discussion

In this paper we have developed a joint Bayesian–frequentist approach for knowledge acquistion (KA) and consensus formation (CF) in populations, about a parameter x that belongs to an opinion space $\mathcal{X}$. We found that KA generally is a more restrictive concept that CF in situations where one parameter value x_0 is correct. Indeed, KA

in a population not only requires that all agents reach consensus, but also that this consensus is about x_0.

The results of this article can be generalized in several ways. First, we have assumed that all agents have access to the same data (before missing out some of them). It is possible to consider scenarios where to some extent agents receive different datasets, also before the missingness mechanism takes place. When this is the case, an agent's motivation to be influenced by agents with different data increases. For instance, if one agent represents an individual with subject expertise, and his dataset is a lot larger than the other agents', he will most likely be influential for many of the other agents.

Second, we have found (see Corollary 3) that a reason for not attaining full knowledge asymptotically when the size N of the data set increases, is biased missing data. This happens, for instance, when agents A_i deliberately suppress or leave out some data points that do not align with their preferences $y_i \in \mathcal{X}$. This potentially makes them less influential since their fraction v_i of retained data points decreases, showing that censoring of information could be detrimental even for influencers. To compensate for missing data, agents could communicate a more peaked version of their posterior distributions in order to be more influential (cf. Remark 3). Another possibility for the agents to be more influential is to fabricate data that align with their own chosen y_i. For instance, some agents might have access to real data, but on top of that fabricate their own data.

Third, it is of interest to let the weigts w_{ji} of influence depend on data and/or information on whether data points are missed or not. For instance, if a new data point D_k arrives and the missingness indicator variables $M_{1k}, \ldots, M_{nk}$ are publically available, it is possible to define the weight of influence of A_j on A_i for data point D_k as $w_{ji} = w_{ji}(M_{jk}, M_{ik})$, so that the influence of A_j on A_i increases when A_i misses out D_k ($M_{ik} = 0$) whereas A_j does not ($M_{jk} = 1$). Another possibility, when agents iterate their beliefs L times, is for the influence of A_j on A_i after l rounds of iterations to be of the form $w_{ji} = w_{ji}(P_{i,l+1}, P_{j,l+1})$ for $l = 0, \ldots, L-1$, with A_i being more influenced by agents with similar posterior distributions as himself. This is analogous to the way influence is defined for the Deffuant model, and for a DeGroot model where individuals are influenced by peers [25].

Fourth, one may analyze the KA process dynamically and find recursive formulas for how the posterior distributions P_i of all agents A_i change when new data points arrive one by one. It is of interest then to formulate stopping rules, when each agent either stops collecting new data or updates from the other agents based on new data, and forms his opinion regarding the value of x_0.

Fifth, if feature vectors are assigned to all individuals of the population, the data set and/or the missingness mechanism of each agent A_i might depend on his feature vector. This feature vector could include age, social background, educational level, and so on.

Sixth, a maximum entropy, quasi-Bayesian approach to KA was introduced in [11]. We believe this approach can be extended from KA of individuals, to KA and CF in populations.

Seventh, our combined CF and KA approach to opinion formation can potentially be applied to multi-agent problem solving by means of Artificial Intelligence (AI) [35]. In this context $\mathcal{X}$ is the set of possible solutions to a given problem, whereas $x_0 \in \mathcal{X}$ is the (unknown) best solution preferred by humans. The n agents represent different deep learning algorithms that propose solutions in terms of posteriors $P_1, \ldots, P_n$. The aligment problem is the challenge to make sure that all these AI algorithms produce solutions P_i that match human interests, corresponding to KA. If, on the other hand, different algorithms A_i have similar, hidden subgoals that surpass human understanding, the result might be CF without KA among these algorithms.

A Proofs

Proof of Theorem 1. We start by performing a second order Taylor expansion of the log posterior distribution of agent A_i, around the point $x_{i\infty}$:

$$\begin{aligned}
\log[P_i(x)] &= \log[P_i(x_{i\infty})] + \psi_i(\boldsymbol{D}; x_{i\infty})(x - x_{i\infty})^T \\
&+ \tfrac{1}{2}(x - x_{i\infty})H_i(\boldsymbol{D}; x_{i\infty})(x - x_{i\infty})^T + o_p(1) \\
&= \log[P_i(x_{i\infty})] + [\bar{\psi}_{i0}(x_{i\infty}) + N\bar{\mu}_i(x_{i\infty})](x - x_{i\infty})^T \\
&+ \{\textstyle\sum_{k=1}^N [\bar{M}_{ik}\psi(D_k; x_{i\infty}) - \bar{\mu}_i(x_{i\infty})]\}(x - x_{i\infty})^T \\
&+ \tfrac{1}{2}(x - x_{i\infty})[\psi'_{i0}(x_{i\infty}) + N\bar{\mu}'_i(x_{i\infty})](x - x_{i\infty})^T \\
&+ \tfrac{1}{2}(x - x_{i\infty})\{H_i(\boldsymbol{D}; x_{i\infty}) - E[H_i(\boldsymbol{D}; x_{i\infty})]\}(x - x_{i\infty})^T + o_p(1) \\
&= \log[P_i(x_{i\infty})] + \{\textstyle\sum_{k=1}^N \bar{M}_{ik}\psi(D_k; x_{i\infty})\}(x - x_{i\infty})^T \\
&+ \tfrac{N}{2}(x - x_{i\infty})\bar{\mu}'_i(x_{i\infty})(x - x_{i\infty})^T + o_p(1),
\end{aligned} \tag{A.1}$$

for all x with $x - x_{i\infty} = O(N^{-1/2})$, making use of (53) and (57) in the second step of (A.1), and (58) and the Law of Large Numbers in the third step. Let $\hat{x}_{iN}$ be the maximum aposteriori estimate of A_i, i.e. the value of x that maximizes $\log[P_i(x)]$ for a data set of size N. It follows from (A.1) that

$$\hat{x}_{iN} = x_{i\infty} - Z_{iN}\bar{\mu}'_i(x_{i\infty})^{-1} + o_p(N^{-1/2}), \tag{A.2}$$

where

$$Z_{iN} = \frac{1}{N}\sum_{k=1}^N \bar{M}_{ik}\psi(D_k; x_{i\infty})$$

is a random vector of length q. Recall that $\boldsymbol{X}_N = (X_{1N}, \ldots, X_{nN})$ is a vector that contains the n agents' random guesses of x_0, in consistency with their posterior distributions. Write

$$\begin{aligned}
\boldsymbol{X}_N &= \hat{\boldsymbol{x}}_N + \boldsymbol{\varepsilon}_N \\
&= \boldsymbol{x}_\infty - \boldsymbol{Z}_N\bar{\mu}'(\boldsymbol{x}_\infty)^{-1} + \boldsymbol{\varepsilon}_N + o_p(N^{-1/2}),
\end{aligned} \tag{A.3}$$

where $\hat{\boldsymbol{x}}_N = (\hat{x}_{1N}, \ldots, \hat{x}_{nN})$ and $\boldsymbol{Z}_N = (Z_{1N}, \ldots, Z_{nN})$ are vectors of length nq, whereas $\boldsymbol{x}_\infty$ and $\bar{\mu}'(\boldsymbol{x}_\infty)$ are defined in (49) and (61) respectively. We deduce from

(55), (60) and the Central Limit Theorem that

$$\sqrt{N}\boldsymbol{Z}_N \xrightarrow{\mathcal{L}} N(0, \bar{J}(\boldsymbol{x}_\infty)) \tag{A.4}$$

as $N \to \infty$ and from (A.1) that

$$\sqrt{N}\boldsymbol{\varepsilon}_N \xrightarrow{\mathcal{L}} N(0, -\bar{\mu}'(\boldsymbol{x}_\infty)) \tag{A.5}$$

as $N \to \infty$. The theorem follows from (A.3), (A.4), and (A.5), and the fact that $\boldsymbol{Z}_N$ and $\boldsymbol{\varepsilon}_N$ are asymptotically independent. The latter claim is a consequence of (47), and that

$$P_i \sim N(\hat{x}_{iN}, \frac{-\bar{\mu}'(\boldsymbol{x}_\infty)^{-1}}{N}), \tag{A.6}$$

according to (A.1), is an increasingly accurate asymptotic approximation of the posterior distribution of A_i. Indeed, since the right-hand side of (A.6) depends on data $\boldsymbol{D}$ only through $\hat{x}_{iN} = \hat{x}_{iN}(\boldsymbol{D})$ for $i = 1, \ldots, N$, assumption (47) implies that $\hat{\boldsymbol{x}}_N$ is asymptotically independent of $\boldsymbol{\varepsilon}_N$, which is equivalent to $\boldsymbol{Z}_N$ being asymptotically independent of $\boldsymbol{\varepsilon}_N$. □

Proof of Corollary 1. This result follows immediately from the joint asymptotic behaviour of the posterior distributions $P_1, \ldots, P_N$ in Theorem 1, and the definition of asymptotic knowledge acquisition in Definition 4. □

Proof of Corollary 2. According (50), (63), and the definition of the degree of consensus $C = C_N$ in (42), we need to prove that

$$\delta_i = \delta(P_i, P) = \delta_{iN} \to \delta_{i\infty} = 0 \text{ as } N \to \infty \text{ for } i = 1, \ldots, n, \tag{A.7}$$

if and only if (64) holds. Since distance measure (43) is used, it follows from (44)-(45), and (48)–(49) that

$$\delta_{i\infty} = (\bar{x}_\infty - x_{i\infty})\Sigma^{-1}(\bar{x}_\infty - x_{i\infty})^T \text{ for } i = 1, \ldots, n, \tag{A.8}$$

where $\bar{x}_\infty = \sum_{j=1}^n x_{j\infty}/n$. Since Σ is a symmetric, positive definite matrix of order q, it is clear from (A.8) that (A.7) is equivalent to (64). □

Proof of Corollary 3. Recall that the scalar functions $\eta_r(x)$, for $r = 1, \ldots, q$, appear in the formula for the exponential family density (4), whereas their derivatives $\eta_r'(x)$ (a row vector of length q) appear in the score function formula (12). It is convenient to introduce the square matrix

$$\eta'(x) = (\eta_1'(x)^T, \ldots, \eta_q'(x)^T)^T \tag{A.9}$$

of order q. Note in particular that $\eta'(x) = I_q$ is the identity matrix of order q when natural parameters are used in the exponential family model (4). In view of (12), it is

possible to rewrite (54) as

$$\mu_j(x) = \xi_j \eta'(x) - v_j B'(x), \tag{A.10}$$

where

$$\xi_j = E[v_j(D)T(D)], \tag{A.11}$$

and $D \sim f(\cdot; x_0)$. Plugging (A.10) into (53) we deduce

$$\bar{\mu}_i(x) = \bar{\xi}_i \eta'(x) - \bar{v}_i B'(x), \tag{A.12}$$

where

$$\bar{\xi}_i = \sum_{j=1}^{n} w_{ji} \xi_j, \tag{A.13}$$

whereas v_j and $\bar{v}_i$ are defined in (68) and (81) respectively. Inserting (66) into (58), and making use of (67) and (69), we find that $x_{i\infty}$ solves the q-dimensional system of equations

$$B'(x_{i\infty})\eta'(x_{i\infty})^{-1} = \frac{\bar{\xi}_i}{\bar{v}_i} = \sum_{j=1}^{n} \tilde{w}_{ji} \frac{\xi_j}{v_j} = \sum_{j=1}^{n} \tilde{w}_{ji} \tilde{E}_j[T(D)]. \tag{A.14}$$

That (A.14) is equivalent to (66) is a consequence of the fact that $B'(x_{i\infty})\eta'(x_{i\infty})^{-1} = E_{x_{i\infty}}[T(D)]$, which follows from (4) by differentiating the equation $\int_{\mathcal{D}} f(z; x)dz = 1$ with respect to x at $x = x_{i\infty}$. □

Verifying the claim of Remark 4 how renormalized weights affect the asymptotic covariance matrix V. Recall that Theorem 1 gives conditions under which the beliefs of agents $A_1, \ldots, A_n$ converge to limiting values $x_{1\infty}, \ldots, x_{n\infty} \in \mathcal{X}$ at rate $\varepsilon_N = 1/\sqrt{N}$ as the number of data points $N \to \infty$, with an asymptotic covariance matrix $V = V_1 + V_2$. Here V_1 and V_2 are square matrices of order nq that correspond to the two terms of the right-hand-side of (59). Suppose the weights $w_{1i}, \ldots, w_{ni}$ of influence on A_i are updated according to (16) for some constant a_i, for $i = 1, \ldots, n$. We will prove that such a renormalization of weights has no effect on V_1, whereas V_2 is modified. To this end, we deduce from (53), (55), and (57) that a renormalization (16) of weights implies

$$\bar{\mu}_i(x) \mapsto a_i \bar{\psi}_i(x), \quad \bar{\mu}_i'(x) \mapsto a_i \bar{\psi}_i'(x), \quad \bar{J}_{ij}(x_i, x_j) \mapsto a_i \bar{J}_{ij}(x_i, x_j) a_j \tag{A.15}$$

for all $1 \leq i, j \leq n$. Let $\bar{A}_i = \mathrm{Diag}(a_i, \ldots, a_i)$ be a diagonal matrix of order q, with all diagonal elements equal to a_i, whereas $\bar{A} = \mathrm{BDiag}(\bar{A}_1, \ldots, \bar{A}_n)$ is a diagonal matrix of order nq, with matrices $\bar{A}_1, \ldots, \bar{A}_n$ along the block diagonal. It follows from (A.15), and the definitions of the matrices $\bar{J}(\boldsymbol{x})$ and $\bar{\mu}'(\boldsymbol{x})$ in (60) and (61), that a renormalization (16) of weights transforms these matrices as

$$\bar{J}(\boldsymbol{x}) \mapsto \bar{A}\bar{J}(\boldsymbol{x})\bar{A}, \quad \bar{\mu}'(\boldsymbol{x}) \mapsto \bar{A}\bar{\mu}'(\boldsymbol{x}). \tag{A.16}$$

From (A.16), and the definitions of V_1 and V_2 on the right-hand-side of (59), we deduce that a renormalization (16) of weights implies that these two matrices transform as

$$V_1 \mapsto V_1, \quad V_2 \mapsto \bar{A}^{-1} V_2, \tag{A.17}$$

as was to be proved.

Proof of Corollary 4. Suppose data is missing completely at random, according to (23). It follows from (53) and (54) that

$$\begin{aligned} \bar{\mu}_i(x_0) &= \textstyle\sum_{j=1}^n w_{ji} E[v_j(D)\psi(D;x_0)] \\ &= \textstyle\sum_{j=1}^n w_{ji} v_j E[\psi(D;x_0)] \\ &= \bar{v}_i E[\psi(D;x_0)] \\ &= 0, \end{aligned}$$

which, in view of (58), implies (79). (Equation (79) can also be derived from (66), the fact that $\sum_j \tilde{w}_{ji} = 1$, and $\tilde{E}_j[T(D)] = E[T(D)] = E_{x_0}[T(D)]$ for MCAR data.) A similar calculation shows that

$$\begin{aligned} \bar{\mu}_i'(x_0) &= \textstyle\sum_{j=1}^n w_{ji} E[v_j(D)\psi'(D;x_0)] \\ &= \bar{v}_i E[\psi'(D;x_0)] \\ &= -\bar{v}_i J^{-1}(x_0). \end{aligned} \tag{A.18}$$

Inserting (A.18) into (61) we find that

$$\bar{\mu}'(x_0) = -\mathrm{Diag}(\bar{v}_1 J^{-1}(x_0), \ldots, \bar{v}_n J^{-1}(x_0)). \tag{A.19}$$

From (56) we deduce that

$$\begin{aligned} J_{lm}(x_0, x_0) &= v_l v_m^{1(l\neq m)} E[\psi(D;x_0)^T \psi(D;x_0)] \\ &= [v_l v_m + v_l(1-v_l)1(l=m)]J(x_0). \end{aligned} \tag{A.20}$$

Inserting (A.20) into (55) we obtain

$$\begin{aligned} \bar{J}_{ij}(x_0, x_0) &= \textstyle\sum_{l,m=1}^n w_{li} w_{mj} [v_l v_m + v_l(1-v_l)1(l=m)]J(x_0) \\ &= [\bar{v}_i \bar{v}_j + \textstyle\sum_{l=1}^n w_{li} w_{lj} v_l(1-v_l)]J(x_0). \end{aligned} \tag{A.21}$$

It follows from (A.19) and (A.21) that block V_{ij} of the covariance matrix V in (59), satisfies

$$\begin{aligned} V_{ij} &= J^{-1}(x_0)\bar{J}_{ij}(x_0,x_0)J^{-1}(x_0)/(\bar{v}_i\bar{v}_j) + 1(i=j)J^{-1}(x_0)/\bar{v}_i \\ &= [\bar{v}_i\bar{v}_j + \textstyle\sum_{l=1}^n w_{li}w_{lj}v_l(1-v_l)]J^{-1}(x_0)/(\bar{v}_i\bar{v}_j) + 1(i=j)J^{-1}(x_0)/\bar{v}_i, \end{aligned} \tag{A.22}$$

which simplifies to (80). □

Proof of Proposition 1. Clearly, the second step of (85) follows as in the derivation (31). In order to prove the first step of (85), we will use induction to prove

$$P_{i,l+1}(x) = \prod_{j=1}^{n} P_{j1}(x)^{w_{ji}^{(l)}} \tag{A.23}$$

for $l = 1, \ldots, L$. The first step of (85) will then follow by taking $l = L$ in (A.23). We start the induction proof by noting that (A.23) holds for $l = 1$, as a consequence of (31). As for the induction step, suppose $2 \leq l \leq L$, and that (A.23) holds for $l-1$. In view of (83), we find that

$$\begin{aligned} P_{i,l+1}(x) &\propto \textstyle\prod_{m=1}^{n} P_{ml}(x)^{w_{mi}} \\ &\propto \textstyle\prod_{m=1}^{n} \left[\prod_{j=1}^{n} P_{j1}(x)^{w_{jm}^{(l-1)}} \right]^{w_{mi}} \\ &\propto \textstyle\prod_{j=1}^{n} P_{j1}(x)^{w_{ji}^{(l)}}, \end{aligned} \tag{A.24}$$

where in the last step we used $w_{ji}^{(l)} = \sum_{m=1}^{n} w_{jm}^{(l-1)} w_{mi}$, which is a consequence of (84). Hence we have showed that (A.23) also holds for l, and this completes the induction proof. □

Proof of Theorem 3. The decomposition (92) of $\mathcal{A}$ into components, and the asymptotic form (93)–(95) of the weight matrix $\boldsymbol{W}^L = (w_{ji}^{(L)})_{j,i=1}^n$ as $L \to \infty$, is based on the asymptotic theory of Markov processes, see for instance [16]. In order to motivate this; notice that (15) implies that $\boldsymbol{\Pi} = \boldsymbol{W}^T = (\pi_{ij})_{i,j=1}^n$ is the transition matrix of a Markov chain, where $\pi_{ij} = w_{ji}$ can be interpreted as the fraction of influence that A_j has on A_i. Alternatively, if agents are chosen randomly, with probabilities equal to their relative influences, then π_{ij} is the probability that A_i picks influencer A_j. Analogously, $\boldsymbol{\Pi}^l = (\pi_{ij}^{(l)})_{i,j=1}^n$ contains influence probabilities L iterations back in time, with $\pi_{ij}^{(l)} = w_{ji}^{(l)}$ the probability that a chain of picked influencers of length l that starts with A_i, ends with A_j. Let $\{I_l\}_{l=1}^{\infty}$ be a time-homogeneous Markov chain with state space $\mathcal{A}$ and transition probabilities $\pi_{ij} = P(I_l = A_j | I_{l-1} = A_i)$ not depending on l. Then

$$P(I_{L+1} = A_j | I_1 = A_i) = \pi_{ij}^{(L)} \to \begin{cases} b_{tj}, & \text{if } A_i \in \mathcal{B}_t \text{ for some } t = 1, \ldots, m, \\ \sum_{t=1}^{m} \alpha_{it} b_{tj}, & \text{if } A_i \in \mathcal{T}, \end{cases}$$

as $L \to \infty$. Consequently, α_{it} is the probability, if the process starts in a transient state A_i ($I_1 = A_i \in \mathcal{T}$), that it ends up in component $\mathcal{B}_t$ ($I_{L+1} \in \mathcal{B}_t$ for large L).

In order to demonstrate limiting consensus as $L \to \infty$ when $m = 1$ (cf. (96)), we make use of (93)–(95) to find that $\boldsymbol{W}^{(\infty)} = (\boldsymbol{b}^T, \ldots, \boldsymbol{b}^T)$, where $\boldsymbol{b} = (b_1, \ldots, b_n)$ is the asymptotic distribution of the only irreducible component $\mathcal{B} = \mathcal{B}_1$. Since this is equivalent to $w_{ji}^{(L)} \to b_j$ as $L \to \infty$, it follows from (85) that $\delta_i^{(L)} = \delta(P_{i,L+1}, P^{(L)}) \to 0$ as $L \to \infty$, where $\delta(\cdot, \cdot)$ is the distance (40) to quantify consensus and $P^{(L)} = \sum_{i=1}^{n} P_{i,L+1}/n$ the average posterior belief of the whole population after L iterations. In view of (41)–(42), this proves (96). □

Declarations

The authors have no financial or non-fiancial competing interests to enclose. Ola Hössjer wrote a first draft manuscript. Then all authors contributed equally in terms of improving it.

Acknowledgement

This paper was written in honor of Svante Janson on the occasion of his 70th birthday.

References

[1] Acemoglu, D., Ozdaglar, A., ParandehGheibi, A.: Spread of (mis)information in social networks. Games and Economic Behavior, 20, 197-227 (2010)

[2] Acemoglu, D., Dahleh, M., Lobel, I., Ozdaglar, A.: Bayesian learning in social networks. Review of Economic Studies 78, 1201–1236 (2011). https://doi.org/10.1093/restud/rdr004

[3] Acemoglu, D., Ozdaglar, A.: Opinion dynamics and learning in social networks. Dyn Games Appl 1, 349 (2011). https://doi.org/10.1007/s13235-010-0004-1

[4] Banerjee, A., Breza, E., Chandrasekhar, A.G., Mobius, M.: Naive learning and uninformed agents. American Economic Review 111(11), 3540–3574 (2021). https://doi.org/10.1257/aer.20181151

[5] Ben-Naim, E., Krapivsky, P.L., Redner, S.: Bifurcations and patterns in compromise processes. Physica D: Nonlinear Phenomena 183 (3–4), 190–204 (2003). https://doi.org/10.1016/S0167-2789(03)00171-4

[6] Bernardo, J.M., Smith, A.F.M.: Bayesian Theory. Wiley (2000)

[7] Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. Reviews of Modern Physics 81, 591–646 (2009). https://doi.org/10.1103/RevModPhys.81.591

[8] DeGroot, M.H: Reaching a consensus. Journal of the American Statistical Association 69, 118–121 (1974). https://doi.org/10.2307/2285509

[9] DeMarco, P.M., Vayanos, D., Zwiebel, J: Persuation bias, social influence, and unidimensional opinions. Q. J. Econ 118(3), 909–968. https://doi.org/10.1162/00335530360698469

[10] Deffuant, G., Neau, D., Amblard, F., Weisbuch, G.: Mixing beliefs among interacting agents. Advances in Complex Systems, 03(01n04), 87–98 (2000). https://doi.org/10.1142/S0219525900000078

[11] Díaz-Pachón, D.A., Gallegos, R., Hössjer, O., Rao, J.S.: Statistical learning does not always entail knowledge. To appear in Bayesian analysis (2025). https://doi.org/10.1214/25-BA1553

[12] Díaz-Pachón, D.A., Hössjer, O., Mathew, C.: (2024). Is it possible to know cosmological fine-tuning? The Astrophysical Journal Supplement Series 271, 56 (2024). https://doi.org/10.3847/1538-4365/ad2c88

[13] Easley, D., Kleinberg, J.: Networks, crowds and markets: Reasoning about a highly connected world. Cambridge University Press (2010)

[14] Genest, C,, Weerahanidi, S., Zidek, J.V.: Aggregating opinions through logarithmic pooling. Theory and Decisions 17, 61–70 (1984)

[15] Ghosal, S., van der Vaart, A.: Fundamentals of Nonparametric Bayesian Inference. Cambridge University Press (2017)

[16] Grimmet, G.R., Stirzaker, D.: Probability and Random Processes, fourth edition. Oxford University Press (2020)

[17] Harris, J.K.: An Introduction to Exponential Random Graph Modeling. SAGE Publications, Inc (2014)

[18] Hirscher, T.: Consensus Formation in the Deffuant Model. Ph.D. Thesis, Division of Mathematics, Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden (2014)

[19] Häggström, O., Hirscher, T.: Further results on consensus formation in the Deffuant model. Electronical Journal of Probability 19, 1–26 (2014). https://doi.org/10.1214/EJP.v19-3116

[20] Hössjer, O., Nyberg, T., Petrovic, S., Sjöberg, F., Öberg, T.: Bayesian distributed detection by sensor networks with missing data. Report 2012:6, Mathematical Statistics, Stockholm University (2012)

[21] Hössjer, O., Díaz-Pachón, D.A., Rao, J.S.: A formal framework for knowledge acquisition: Going beyond machine learning. Entropy 24, 1469 (2022). https://doi.org/10.103390/e24101469

[22] Hössjer, O., Díaz-Pachón, D. Chen, Z., Rao, S.: An information theoretic approach to prevalence estimation and missing data. IEEE Transactions on Information Theory 70(5), 3567–3582 (2024). https:/doi.org/10.1109/TIT.2023.3327399.

[23] Ichikawa, J.J., Steup, M.: The Analysis of Knowledge. The Stanford Encyclopedia of Philosophy, Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA (2018)

[24] Jackson, M.O: Social and Econmic Networks. Princeton University Press (2008)

[25] Krause, U: A discrete nonlinear and non autonomous model of consensus formation. In: Elaydi, S., Ladas, G., Popenda, J., Rakowski, J. (eds.) Communications in difference equations, pp. 227–236, Gordon and Breach Science Publishers, Amsterdam (2000)

[26] Lanchier, N.: Stochastic Interacting Systems in Life and Social Sciences. De Gruyter (2024)

[27] Lehmann, E.L., Casella, G.: Theory of Point Estimation. Springer (1998)

[28] Ligget, T.M: Interacting Particle Systems. Springer, New York (1985)

[29] Little, L., Rubin, D.: Statistical Analysis of Missing Data, 3rd edition. Wiley Series in Probability and Statistics (2019)

[30] Lusher, D., Koskinen, J., Robins, G., (eds.): Exponential Random Graph Models for Social Networks: Theory, Methods and Applications (Structural Analysis in the Social Sciences). Cambridge University Press, Cambridge (2012)

[31] Lyons, R., Peres, Y.: Probability on Trees and Networks. Cambridge University Press, Cambridge (2016)

[32] Meng, X.F., Van Gorder, R.A., Porter, M.A.: Opinion formation and distribution in a bounded-confidence model on various networks. Physical Review E 97(2), 022312 (2018). https://doi.org/10.1103/PhysRevE.97.022312

[33] Molavi, P., Tahbaz-Salehi, A., Jadbabaie, A.: A theory of non-Bayesian social learning. Econometrica 86(2), 445–490 (2018). https:/doi.org/10.3982/ECTA14613

[34] Nagarajan, R., Scutari, M., Lébre, S.: Bayesian Networks in R with Applications in Systems Biology, Springer (2013)

[35] Ngo, L., Chan, L., Minderman, S.: The alignment problem from a deep learning perspective. arXiv:2209.00626v8 (2025)

[36] Proskurnikov, A.V., Tempo, R.; A tutorial on modeling and analysis of dynamic social networks. Part I. Annual Reviews in Control 43, 65–79 (2017). https://doi.org/10.1016/j.arcontrol.2017.03.002

[37] Olfati-Saber, R., Fax, J.A., Murray, R.M.: Consensus and cooperation in networked muli-agent systems. Proceedings of the IEEE 95(1), 215–233 (2007)

[38] Saligrama, V., Alanyali, M., Savas, O.: Distributed detection in sensor networks with packet losses and finite capacity links. IEEE Transactions on Signal Processing 54(11), 4118–4132 (2006)

[39] Thorvaldsen, S., Hössjer, O.: Use of directed quasi-metric distances for quantifying the information of gene families. BioSystems 243, 105526 (2024). https://doi.org/10.1016/j.biosystems.2024.105256.