Multi-model and network inference based on ensemble estimates: avoiding the madness of crowds

Michael P. H. Stumpf¹,²

¹School of BioSciences and School of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010, Australia
²Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, London SW7 2AZ, UK

Recent progress in theoretical systems biology, applied mathematics and computational statistics allows us to compare the performance of different candidate models at describing a particular biological system quantitatively. Model selection has been applied with great success to problems where a small number—typically less than 10—of models are compared, but recent studies have started to consider thousands and even millions of candidate models. Often, however, we are left with sets of models that are compatible with the data, and then we can use ensembles of models to make predictions. These ensembles can have very desirable characteristics, but as I show here are not guaranteed to improve on individual estimators or predictors. I will show in the cases of model selection and network inference when we can trust ensembles, and when we should be cautious. The analyses suggest that the careful construction of an ensemble—choosing good predictors—is of paramount importance, more than had perhaps been realized before: merely adding different methods does not suffice. The success of ensemble network inference methods is also shown to rest on their ability to suppress false-positive results. A Jupyter notebook which allows carrying out an assessment of ensemble estimators is provided.

1. Introduction

In physics, simple and elegant symmetry relationships have often led the way to theoretical models [1]. Most importantly Emmy Noether’s theorem has been pivotal in establishing the correspondence between continuous symmetries and conservation laws in physics [2]; it has been instrumental in the derivation of physical laws of nature. Biology has not been able to draw on such fundamental principles [3], to a large degree because most processes are intrinsically dissipative (meaning energy is produced and consumed) and hence the conditions where Noether’s theorem holds simply do not apply. Instead, biological models have often had a heuristic element or described the rates of change in (often macroscopic) observables (e.g. species of plants or animals, or molecular species).

Writing down the set of equations is an important starting point in modeling as it forces us to express our assumptions in a precise form. Which form to choose is, however, not unambiguously obvious. Instead, we often rely on data to decide between the different options. Statistical model selection [4] provides the tools to make such decisions, balancing the ability of a model to fit the data with the model’s complexity. As larger and larger models, even models for whole cells [5–8], are being considered model selection problems will presumably become the norm, especially when models are constructed exhaustively or automatically [9–16].
In some, probably many, situations model selection will not be able to decide on a single best model. Instead, many models may have comparable support. In such a situation, we may then base analysis or predictions on the ensemble of models that are supported by the data [17]. Each model's contribution to the prediction etc. is weighted by the relative support it has received. Such estimates or predictions based on ensembles have been referred to as exploiting the ‘wisdom of crowds’ [18]. This refers to the notion that groups of individuals/models are more likely to be better than those based on a single individual/model. This concept, however, also relates to much earlier work, Charles Mackay’s ‘Extraordinary Popular Delusions and the Madness of Crowds’ [19], a 19th century account of how popular opinion can support quite extraordinary and plainly wrong opinions and concepts.

Ensemble estimators have a surprisingly long history, outlined in [20], and aspects such as bagging, boosting and stacking [21] are firmly established in the statistical learning literature; see, for example, [22] for a comprehensive treatment. There has been interest in evolutionary biology [23]; and following [18], there have been further developments in systems and computational biology, e.g. [24,25]. But in the context of network inference, combining different network reconstruction methods has sometimes been viewed as necessarily optimal [26]. Below, we show that this is not automatically the case. In turn, I will show that model averaging and ensemble estimation are susceptible to poorly defined sets of candidate models; and that the behaviour of ensemble approaches to network reconstruction depends strongly on the composition of the ensemble. For very good ensembles, the advantage comes mainly from reducing the number of false-positive edges. Both problems share a dependence on the quality of the ensemble, and we map out and quantify these influences; we also provide self-contained Julia code for further in silico experimentation and analysis of ensemble prediction methods.

2. Model selection and multi-model inference

We assume that we have a universe of models, \( \mathcal{M} \),

\[ \mathcal{M} = \{ M_1, M_2, \ldots, M_N \} \]

that are potential mechanisms, by which some data, \( D \), have been generated. For simplicity, we consider a finite number of models, \( N \). Furthermore, for each model, \( M_i \), we assume that we have a data generating function, \( f_i(\theta) \), parametrized by a parameter vector \( \theta_i \) which is chosen from some suitable continuous parameter space,

\[ \theta_i \in \Theta_i \subseteq \mathbb{R}^n. \]

Coping with the size of the parameter space is one of the essential challenges of parameter estimation and model selection.

We start from a Bayesian framework [27], where we seek to determine the posterior distribution over parameters,

\[ \Pr(\theta_i | D) = \frac{\Pr(D | \theta_i) \pi(\theta_i)}{\Pr(D)} , \]

(2.1)

where \( \Pr(D | \theta_i) \) is the likelihood, \( \pi(\theta_i) \) the prior over the parameters for model \( M_i \), and \( \Pr(D) = \int \Pr(D | \theta) \pi(\theta) d\theta \) (here, we make the dependence on the choice of model explicit through an index) is known as the evidence.

In the Bayesian framework model selection is a (relatively) straightforward extension, and the model posterior is given by

\[ \Pr(M_i | D) = \frac{\Pr(D | M_i) \pi(M_i)}{\Pr(D)} = \int \Pr(D | \theta) \pi(\theta) d\theta \pi(M_i) \]

(2.2)

where analogously to equation (2.1), we have the model posterior, \( \Pr(M_i | D) \), and model prior, \( \pi(M_i) \). The denominator terms in equations (2.1) and (2.2) are notoriously hard to evaluate for all but the simplest cases, and a large amount of ingenuity and work has been invested into computational schemes [27, 28], including Markov chain Monte Carlo, sequential Monte Carlo and related approaches. Often even the likelihood is prohibitively expensive to evaluate and so-called approximate Bayesian computation schemes have been devised to make Bayesian statistical inference possible [29].

Alternatives to the Bayesian framework, such as likelihood-based inference and optimization of cost functions [30], result in point estimates for the parameters, e.g. the value of \( \theta^* \) that maximizes the probability of observing the data,

\[ \theta_0^* = \arg\max_{\theta} \Pr(D | \theta^*). \]

(2.3)

Similarly, we can determine the maximum a posteriori estimate by finding the mode of the posterior distribution [27],

\[ \theta_B^* = \arg\max_{\theta} \Pr(\theta | D). \]

(2.4)

Compared to analysis of the posterior distribution, such local estimates lose a lot of relevant information, but some characteristics can still be recovered by considering the local curvature of the likelihood, i.e. the information matrix, or cost-function surface around the (local) extremum identified in this manner [31,32].

Model selection frameworks that are based on likelihood inference rely on criteria to find the optimal model among a set of candidate models. The Akaike information criterion (AIC) [4,33] for model \( M_i \) is given by

\[ \text{AIC}_i = -2 \log(\Pr(D | \hat{\theta}, M_i)) + 2n_i, \]

(2.5)

with \( \hat{\theta} \) given by equation (2.3), and \( n_i \) the number of parameters of model \( M_i \). The AIC is probably the most widely used model selection criterion, despite the fact that it is biased in favour of overly complicated models as the amount of available data increases. The Bayesian information criterion does not suffer in the same way; it is defined as

\[ \text{BIC}_i = -2 \log(\Pr(D | \theta, M_i)) + n_i \log(n), \]

(2.6)

with \( n \) representing the size of the data or number of samples. Several other information criteria exist (discussed e.g. in [4,34]), but they all share in common the purpose of balancing model complexity with model fit. The BIC can be derived as an approximation to the full Bayesian model posterior probability, which achieves this balance implicitly.

If model selection cannot pick out a clear winner, then either (i) further analysis should be used to design better, more informative experiments that can discriminate among these models [35–37], or (ii) these models should be considered as an ensemble [4,25]. The former has definite attractions as it can support quite extraordinary and plainly wrong opinions. Crowds’ views lead to an increase in our understanding if we can discard some of the models/mechanistic hypotheses.

The latter approach, basing analysis and especially predictions on an ensemble of models has become a popular concept.
in machine learning. Most notably, in the context of biological network inference the ‘wisdom of crowds’ concept [18] has been important in popularizing inference based on several models. Here, we are considering model averaging where contributions from different models may be weighted by their relative explanatory performance. In the Bayesian framework, we can use the posterior probability directly. In the context of an information criterion \( I \), for model \( i \), we define [4]

\[
\Delta_i = I_i - \text{argmin}_i I_i,
\]

and then determine the model weight as

\[
w_i = \frac{\exp(-\Delta_i)}{\sum_{i=1}^N \exp(-\Delta_i)},
\]

(2.7)

The model weights (e.g. the Akaike weight if \( I_i \) is the AIC) provide the relative probability of model \( M_i \) to be the correct model conditional on the data \( D \) and the set of models, \( M \), considered. Model averaging in this framework can serve as a variance reduction technique [4,21].

### 3. Statistical physics of model selection and ensemble estimation

In order to simplify the discussion, we define a relationship between the model probability (always implicitly understood to be either a posterior or relative model probability), \( p_i \), and a cost or energy, \( \epsilon_i \), [8] as

\[
p_i = \frac{\exp(-\beta \epsilon_i)}{Z},
\]

(3.1)

with the normalization constant \( Z \) (the partition function) given by

\[
Z = \sum_{i=1}^N \exp(-\beta \epsilon_i).
\]

(3.2)

With this in hand, we can straightforwardly consider different model selection/averaging frameworks in a similar manner.

In general, the true data-generating model (a natural system) will not be represented perfectly by any of the models in \( M \); we denote this true model by \( \hat{M}_i \). But if we are interested in finding out whether \( \hat{M} \) has a certain characteristic \( \phi \) we would obtain this from the appropriate ensemble average

\[
\Pr(\phi \in \hat{M}) = \sum_{i=1}^N p_i \mathbb{I}(\phi \text{ is part of } M_i),
\]

(3.2)

where \( \mathbb{I}(x) \) is the conventional indicator function, i.e.

\[
\mathbb{I}(x) = \begin{cases} 
1 & \text{if } x \text{ is true}, \\
0 & \text{otherwise}.
\end{cases}
\]

Equation (3.2) is based on three, potentially strong assumptions.

1. The model universe, \( M \), is ‘complete’ in the sense that we expect \( M \) to contain any model, \( M_i \), which we might expect to be a reasonable description of \( N \) (always remembering that \( N \notin M \)).
2. It is decidable if \( \phi \) is part of \( M_i \), \( \forall M_i \in M \).
3. \( \phi \) has played no part in the construction of the model probabilities, \( p_i \), equation (3.1).

The first assumption is arguably the strongest assumption. In principle, we can update \( M 
\rightarrow M' \) by adding additional models in light of data; specifying new priors \( \pi' \) for the models \( M_i' \in M' \) will require some care. One important condition that \( M \) (and \( M' \)) must fulfil is that models must not appear more than once in the set. This is important to keep in mind as we are increasingly relying on automated generation or exhaustive enumeration of models.

With fixed \( D \) and \( M \) equation (3.2) is, however, a good approach for ensemble-based prediction and estimation. It also encompasses Bayesian model averaging and multi-model inference based on information criteria [4].

#### 3.1. The effective model space

We first analyse a simple case where all models in our universe have associated costs drawn from a suitably well behaved probability distribution, \( \epsilon_j(q) \) [8], with positive support and associated density, \( f_j(\epsilon) \), over the model energies, \( \epsilon_j \) such that

\[
\epsilon_j \sim \epsilon(q).
\]

Because \( \epsilon_j \) is a random variable, the relative weight, \( w_i = \exp(-\epsilon_j) \), will also be a random variable, and we can obtain the probability density function, \( f(\omega) \), via change of variables as

\[
f(\omega) = \frac{q(-\log(\omega))/|\omega|}{\omega I(\cdot)}.
\]

(3.4)

For different choices of \( q(\epsilon) \), we can now investigate the distribution over model weights. For example, if \( \epsilon_j \sim \text{Gamma}(\alpha, \theta) \) (where \( \alpha \) and \( \theta \) denote the shape and scale parameters, respectively), then

\[
f(\omega) = \frac{\log(\omega)^{\alpha-1} \theta^\alpha}{\omega^{\alpha} I(\cdot)} \text{,}
\]

(3.5)

with \( I(\cdot) \) the Gamma function. The Gamma distribution is a flexible and generic distribution and is chosen for its generality rather than any particular property and our discussion here does not depend on its specifics. Some representative distributions over \( \epsilon \) and the corresponding distributions over \( \omega \) are shown in figure 1a,b.

The distribution over model costs, \( \epsilon \), affects the distributions over model weights, \( \omega \). This is important to realize when deciding on how to triage model sets for further analysis and prediction [38]; generally, inference based on all models weighted by \( \omega \) is neither practical nor desirable, as many models with low weight will mask the information available in the good models. If, for example, we only include models with \( \omega \in [0.9, 1.0] \) then the average model cost

\[
e_{\omega \geq \omega} = \int_{1}^{0} -\log(\omega) f(\omega) \, d\omega
\]

(because \( \epsilon = -\log(\omega) \) for these sets will be 0.005 (blue), 0.0015 (green), \( 6 \times 10^{-4} \) (purple) and \( 9 \times 10^{-5} \) (orange), see figure 1c. The (unknown) distribution over costs can affect multi-model inference quite profoundly. But for model universes that are enriched for good models (i.e. many models \( M_i \) with low values of \( \epsilon_j \)) selecting a subset of models based on even a fairly conservative threshold for the model weights \( \omega_i \) can result in a sufficiently large model sets for further prediction.

#### 3.2. A simple test case for multi-model inference

Here, we study a very simplistic scenario in which we have three types of models, borrowing and adapting Box’s [39] terminology:

**Useful Models** which capture important aspects of \( N \) and which have an associated cost \( \epsilon_1 \).
Useless Models which are qualitatively and quantitative poor descriptions of reality and have an associated cost $\epsilon_2 \gg \epsilon_1$.

Nuisance Models which are qualitatively different from reality, but which can quantitatively capture aspects of $\hat{\eta}$ by chance; their costs are random variables, $\eta$.

Nuisance models are here assumed to be models where the quantitative agreement with data is unrelated to the data generating mechanism $\hat{\eta}$. Purely machine-learning-based models are one way in which we can realize such nuisance generating mechanism [40]; for small datasets $D$, poorly designed experiments, or simply lack of prior knowledge, there are many ways in which model fit may only be a poor reflection of reality, and this can also give rise to nuisance models [41].

For concreteness let these three different model classes have sizes $v_1$, $v_2$ and $v_3 = N - v_1 - v_2$, and assume that $\eta \sim U[0,1]$, i.e. nuisance models are at worst as bad as useless models. Then the number of nuisance models that have lower associated costs than the useful models is given by

$$\frac{\epsilon_1}{\epsilon_2} v_3.$$ 

The relative influence of nuisance models can be studied by contrasting three features, $\phi_1$, $\phi_2$, and $\phi_3$, with the following properties:

$\phi_1$: equally represented with frequency $\xi$ among models of all classes.
$\phi_2$: only represented among useful models.
$\phi_3$: only represented among ‘nuisance’ models.

With equation (3.2) we can obtain $\text{Pr}(\phi_i \in \mathcal{N})$ for any property with frequencies $\xi_i$ in the classes $i = 1, 2, 3$,

$$\text{Pr}(\phi_i \in \mathcal{N}) = \frac{v_1 \xi_1 e^{-\epsilon_1} + v_2 \xi_2 e^{-\epsilon_2} + (v_3 \xi_3 / \epsilon_2) \int_0^{\epsilon_2} e^{-\epsilon_2} d\eta}{v_1 e^{-\epsilon_1} + v_2 e^{-\epsilon_2} + (v_3 / \epsilon_2) \int_0^{\epsilon_2} e^{-\epsilon_2} d\eta} = \frac{v_1 \xi_1 e^{-\epsilon_1} + v_2 \xi_2 e^{-\epsilon_2} + (v_3 \xi_3 / \epsilon_2) (1 - e^{-\epsilon_2})}{v_1 e^{-\epsilon_1} + v_2 e^{-\epsilon_2} + (v_3 / \epsilon_2) (1 - e^{-\epsilon_2})}. \tag{3.6}$$

First, for $\phi_1$, we trivially obtain

$$\text{Pr}(\phi_1 \in \mathcal{N}) = \xi.$$ \tag{3.7}

For the more interesting probability for $\phi_2$ under the model averaging framework, we obtain

$$\text{Pr}(\phi_2 \in \mathcal{N}) = \frac{v_1 \xi_1 e^{-\epsilon_1}}{v_1 e^{-\epsilon_1} + v_2 e^{-\epsilon_2} + (v_3 / \epsilon_2) (1 - e^{-\epsilon_2})}, \tag{3.8}$$

and finally, for a characteristic shared by and confined to the set of nuisance models, we obtain

$$\text{Pr}(\phi_3 \in \mathcal{N}) = \frac{(v_3 / \epsilon_2) (1 - e^{-\epsilon_2})}{v_1 e^{-\epsilon_1} + v_2 e^{-\epsilon_2} + (v_3 / \epsilon_2) (1 - e^{-\epsilon_2})}. \tag{3.9}$$

From equations (3.8) and (3.9), we see that multi-model averaging is prone to propagate wrong results if nuisance models are frequent and some receive good quantitative support (i.e. low model costs, $\epsilon$). Equally worrying, the same scenario can make it hard for true aspects of $\eta$ to receive sufficient support through equation (3.8) if there are many nuisance and useless models included in the data.

To illustrate this further we can consider the case where $\xi_2 = 1$ and $\xi_3 = 1$ (meaning every useful model exhibits characteristic $\phi_2$, and every nuisance model characteristic $\phi_3$) and ask when is $\text{Pr}(\phi_2 \in \mathcal{N}) > \text{Pr}(\phi_3 \in \mathcal{N})$? We obtain

$$v_3 > \epsilon_2 v_1 \frac{e^{-\epsilon_1}}{1 - e^{-\epsilon_1}} > \epsilon_2 v_1 e^{-\epsilon_1}. \tag{3.10}$$

Thus, if useful models are sufficiently rare in the model set (say $v_1 < 0.1$) the nuisance models’ characteristics will have high weight in the ensemble average; see also figure 2. None of the parameters $\nu_1$, $\nu_2$, $\nu_3$, $\epsilon_1$, $\epsilon_2$ are, of course, known, and we cannot know which class a model belongs to $\text{a priori}$. Thus model averaging is not a panacea and requires careful consideration of which models are included in the prediction set.
4. Ensemble estimation and network inference

Network inference can also benefit from ensemble methods [18,24,42] but here, too, potential pitfalls arise. We are considering directed networks, with $V$ nodes, and $L$ edges; the adjacency matrix, $A$, is a convenient way to specify such networks, if we indicate the presences and absences of edges, by one and zero, respectively, in its entries,

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge from } i \text{ to } j, \\ 0 & \text{otherwise.} \end{cases}$$

In network inference, we seek to determine whether the data $D$ suggest statistical relationships among pairs of nodes $v_i$, $v_j$ that may indicate a functional relationship, represented by an edge connecting the nodes. We consider $k$ different algorithms, $O_{i\kappa} = 1, \ldots , k$, which predict the presence $A_{ij}^{(\kappa)} = 1$ or absence $A_{ij}^{(\kappa)} = 0$ of a link from $i$ to $j$. If we have for the false positive and false-negative probabilities for method $\kappa$,

$$s_\kappa = \Pr(1 \mid \text{edge is not part of the true network})$$

and

$$t_\kappa = \Pr(0 \mid \text{edge is part of the true network}),$$

we can assess how beneficial the ensemble estimators are for the quality of the inferred networks.

4.1. Ensembles of identical estimators

The simplest case, which is already instructive as a baseline, is where all methods have identical performance, $s_\kappa = s$ and $t_\kappa = t$, $\forall \kappa = 1, \ldots , k$. If the performance of the inference methods is statistically independent, then the number of agreeing inference methods is a binomial random variable. If we set a threshold on the minimum number of concordances among inference methods, $k_0$, we observe for the overall probability of scoring a true edge from the ensemble,

$$\tilde{\Pr}(1 \mid \text{edge is in } R) = \sum_{\kappa = k_0}^{k} ks(1 - t)^{k - \kappa},$$

while the probability of a false negative is

$$\Pr(1 \mid \text{edge is not in } R) = \sum_{\kappa = k_0}^{k} k_0 s(1 - s)^{k - \kappa}. \quad (4.4)$$

To illustrate the outcome of such a simple estimation procedure, we assume a network loosely based on expected Homo sapiens network sizes [43] (22,000 nodes and 750,000 interactions). In figure 3, we consider 10 ensembles and base the ensemble estimator on the minimum number, $k_0$, of methods that have to predict an edge. If $k_0$ is too small, then there will be too many positives as is the case here; a majority-vote rule, i.e. here $k_0 = 6$, would do an acceptable job in terms of precision, recall and F1 statistics [21], but this does depend on the false positive ratio in particular (as biological networks are sparse), as well as the size of the ensemble of inference algorithms, $O = \{O_1, O_2, \ldots , O_k\}$.

4.2. Ensemble estimators can be worse than individual estimators

We are interested in ensemble estimators because we know that individual estimators are far from perfect. But ensembles are not guaranteed to always improve on individual estimators. We compare an ensemble of equally well performing estimators with a single estimator. The ensemble false negative probability, $T$, is given by

$$T = \sum_{\kappa = k_0}^{k} \frac{k}{k_0} t^\kappa (1 - t)^{k - \kappa}$$

$$= \frac{k}{k_0} \binom{k}{k_0} (1 - t)^{k - k_0} \text{F}_1 \left( 1, k_0 - k; k_0 + 1; \frac{t}{t - 1} \right), \quad (4.5)$$

where $\text{F}_1$ is the hypergeometric function [47] (see appendix B for an approximation for small arguments). From this, we can determine when the ensemble false negative rate, $T$, will be greater than $t$, i.e.

$$t < \frac{k}{k_0} \binom{k}{k_0} (1 - t)^{k - k_0} \text{F}_1 \left( 1, k_0 - k; k_0 + 1; \frac{t}{t - 1} \right).$$

Figure 2. For $t_1 = 0.1$, we indicate the areas in the $\epsilon_2$, $\nu_1$ space where the contribution to ensemble estimators coming from the nuisance models exceeds that coming from the useful models (black). (a) The frequency of useful models is $\nu_1 = 0.01$; in (b), we set $\nu_1 = 0.1$. 

(a) $\nu_1 = 0.01$

(b) $\nu_1 = 0.1$
Equally, we obtain for the ensemble false positive rate, $S$,

$$
S = \frac{k}{k_0} F^k (1 - s)^{k_0-k} _2F_1\left(1, k_0-k; k_0+1; \frac{s}{s-1}\right). \quad (4.6)
$$

For low thresholds, $k_0$, the ensemble error rate can be greater than that of the individual estimator; this is because the stringency of the ensemble prediction is then reduced, as the cumulative probability of a sufficiently small number of estimators to ‘accidentally’ agree is greater than the error rate of the individual estimator. We show this for two false negative rates, $t = 0.1$ and $t = 0.05$, in figure 4a,b.
4.3. Heterogeneous ensembles of network inference methods

We next focus on the case of a small set of predictors, \( k = 10 \), and two classes of methods: a set of good methods with error rates \( t_1 \) and \( s_1 \), and a set of poor methods with error rates \( t_2 > t_1 \) and \( s_2 > s_1 \). In figure 5, we show, again for a case modelled on a likely human gene regulation network, the likely true- and false-positive predictions arising from ensembles with different numbers of good versus poor edge prediction methods. The basic lesson is that good methods have to outnumber bad methods; otherwise, especially the precision will suffer. Here, we have chosen a simple majority-vote criterion. To bring precision and F1 statistic up to a satisfying level (say in excess of 0.7) requires essentially purging the ensemble of the weakly performing methods (i.e. \( k_1 \approx 8 \)). This only points to the extent to which poor methods can compromise the performance of

---

**Figure 5.** Illustration of the performance of ensembles of network inference methods with false positive and false negative probabilities \( s = t = 0.1 \) for the good predictors and with \( s = t = 0.15 \) (a–c) and \( s = t = 0.25 \) (d–f). Real biological networks are sparse unless false-positives are controlled in the ensemble; once false-positives are over-controlled (by demanding a larger number of methods to score an edge), the recall deteriorates. This is reflected in (a,b,d,e), showing the numbers of true and false positives (we generated 1000 random inferred networks). (c,f) The precision, recall and F1 statistics [21] (see also appendix A) as a function of the minimum number of methods that need to positively score an edge. Increasing the false-positive and false-negative error rates from 0.1 to 0.25 for the poor estimators results in marked deterioration of the ensembles. And even a small number of the bad predictor can profoundly affect the ensemble performance.
ensemble estimators (and the accompanying Jupyter notebook can be used to explore this further).

In sparse networks especially, poor estimators will result in inflation of false-positive results, and lead to overall poor performance: in a directed network there are \( N \times (N - 1) \) possible interactions, but this is still vastly greater than the number of existing edges. For example, in the case of the human network we would have of the order of \( 4.8 \times 10^9 \) potential interactions of which only some 750 000 are expected to exist. So even for \( s_2 = 0.5 \) there would be about 470 000 cases where ten such methods would agree and score an edge, and \( 1.8 \times 10^8 \) false predictions, if a simple majority vote rule were applied.

5. Discussion

We have shown that ensemble estimators are not as robust as has sometimes been claimed [26], or incorrectly surmised from the success of community average results in the DREAM competition [18]; there, of course, it had already been shown that certain, carefully selected, subsets of estimators give better results than others [24].

For the analysis of multi-model inference from mechanistic models, we can distill two points: (i) ensembles of mechanistic models that are reasonably defined [48] (i.e. their construction incorporates any available mechanistic insights; duplicate models are avoided; the model is predictive and can be used to generate data that can be compared with real observations/data) can be combined with the aid of model selection criteria or Bayesian posterior model probabilities with relative ease and safety; (ii) the inclusion of ‘nuisance models’ can hamper ensemble approaches if they come to predominate the model universe \( M \). Such situations could become more likely as model spaces are explored exhaustively [16] or automatically [12]. Because of the formalism connecting different model selection criteria, equation (3.1), these are general results, and do not depend on the particular model averaging procedure chosen (as also clear from the analysis in [20]). So, in essence, the construction of the models in \( M \) [3,49] will determine the robustness of model averaging or ensemble approaches to prediction and analysis. Little is to be gained by increasing the size of \( M \) beyond the (already quite large) set of reasonable models.

In the context of network inference, the situation is similar. We find that the poor performance of some methods can drag down the performance of an ensemble estimator or network predictor. So like in the construction of the model universe \( M \) before, the make-up of the ensemble of network inference methods, \( O_{nk} = 1, \ldots, k \), does matter considerably (as was also found in the empirical study of [18]). Majority vote will typically be a sound criterion for a set of reasonable estimators, though not necessarily optimal from the perspective of precision and F1 measures. This is because biological networks are sparse and false-positives will predominate the inferred networks unless they are carefully controlled. For a set of statistically similar powerful inference methods, a conservative criterion for scoring edges will improve on the overall performance of the individual estimators, however.

The problem of network inference has long been known to be challenging. One reason for this (in addition to the large-scale testing problem) is that we do not have a fair way to score and compare the performance of different network inference approaches. The most promising existing approaches are typically computationally expensive and rigorous in silico assessment of performance as well as the factors influencing performance is often seen as computationally prohibitively expensive. There is also a danger of biasing simulated data in favour of a given method and the DREAM competition has aimed—with some success we feel—to avoid this, and other approaches have followed suit. Clearly, more effort in this domain is needed and computational cost should not preclude rigorous assessment [50]. This situation is mirrored in other areas of genomics and systems biology, e.g. pseudo-temporal ordering approaches have until recently [51] rarely been rigorously tested. But what is also needed are approaches which allow us to assign confidence to inferred networks, or, more specifically, predicted interactions without recourse to a gold-standard [52]. Here, measures based on biological expectations/knowledge offer promising routes for filtering out poor methods [53] (see figure 6).

Figure 6. Illustration of different predictors which capture different aspects of the data. If predictors \( O_1, O_2 \) and \( O_3 \) have particular performance preferences for certain types of interactions then combining them may improve the overall estimate. But this depends crucially on the false-positive rate. In an ideal scenario, we would also be able to exploit the individual strengths of the predictors to reduce the ensemble false negative rate; the mathematical formalisms exist [20,22,54,55], but we need to be able to quantify the individual predictors’ behaviours better.

One of the potential initial attractions of using a panel of network inference algorithms is that different methods may capture different aspects of the data and in concert may provide a more complete representation of a ‘true’ network of functional relationships among the genes in an organism under a given scenario. While appealing this notion needs to be viewed with caution [56]. Combining the most powerful methods by leveraging their individual strengths is possible in principle [20,22,54,55], but requires us to characterize each method \( O_i \) reliably and independently.

6. Conclusion

In summary, unless we know the constituency of the model universe, \( M \), or the ensemble of predictors, \( O_\star \), we have limited ways of telling whether we are dealing with a madness of crowds or a wisdom of crowds scenario. However, the present analysis shows that ensemble procedures will be robust as long as the ensembles are carefully constructed. In the context
of biological network inference, reduction in false-positives is the primary cause of their success. Without a robust and transferrable way of assessing the strengths and weaknesses of different methods, we cannot (yet) use tools from decision theory that pool these strengths for diverse and heterogeneous ensembles [54, 55]. Currently, the best advice, in light of the analysis carried out here, is to be ruthless in weeding out poorly performing methods for network inference, or models with low weight for multi-model averaging. So there is no need, for example, to include correlation networks, even though they are cheap to calculate: their performance is simply too poor to warrant inclusion in an ensemble. Quality is more important than quantity.

Data accessibility. A Jupyter Notebook containing the Julia code to reproduce all the computational results here, and to explore the effects of the madness of crowds in network inference and model averaging is available at https://github.com/MichaelPHStumpf/Madness-Of-Crowds.

Competing interests. I declare I have no competing interest.

Funding. This work has been supported by the University of Melbourne Driving Research Momentum Program. Funding has been received from the Volkswagen Stiftung through a ‘Life?’ Program grant.

Appendix A. Assessing the performance of network inference methods

Real biological networks are sparse [57]; this means that a predictor which scores each candidate edge to be absent can have high performance if the number of false-positives is heavily influencing how we quantify performance of a network inference method. We therefore focus on precision and recall and a derived statistic, the F1 statistic [21, 57]. We denote the numbers of true and false positive inferences (i.e. scored edges in the context of network inference) by TP and FP, and the true and false negatives by TN and FN.

Then the precision, P, is given by

\[ P = \frac{TP}{TP + FP} \]

and the recall by

\[ R = \frac{TP}{TP + FN} \]

The F1 statistic is given by

\[ F1 = 2 \cdot \frac{P \cdot R}{P + R} \]

These statistics are confined to the [0, 1] range, with larger values indicating better performance, and the F1 statistic in particular becomes maximal at the point where the curves for P and R intersect.

Appendix B. Approximating the hypergeometric function

The hypergeometric function [47], 3F1, appearing in equation (4.5) can be unwieldy to evaluate for large k and k_0. For sufficiently small values of t, we can Taylor-expand it around t = 0 and then obtain (if we restrict the expansion of the hypergeometric function to third order)

\[ T = \frac{k}{k_0} (1 - t)^{k - k_0} \left( 1 - \frac{k - k_0}{k + 1} \times \left( 1 + \frac{t(2k_0 - k - 3)}{(k_0 + 2)} \right) \right) \]

From this, we can also determine when the ensemble false-negative rate \( t \) will be greater than 1 by solving

\[ t > \frac{k}{k_0} (1 - t)^{k - k_0} \]

References

1. Thorne KS, Blandford RD. 2017 Modern classical physics: optics, fluids, plasmas, elasticity, relativity, and statistical physics. Princeton, NJ: Princeton University Press.
2. Neuenschwander DE. 2017 Emmy Noether’s wonderful theorem. Baltimore, MD: JHU Press.
3. May RM. 2004 Uses and abuses of mathematics in biology. *Science* 303, 790–793. (doi:10.1126/science.1094442)
4. Burnham KP, Anderson DR. 2013 *Model selection and inference: a practical information-theoretic approach*. New York, NY: Springer Science & Business Media.
5. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JJ, Covert MW. 2012 A whole-cell computational model predicts phenotype from genotype. *Cell* 150, 389–401. (doi:10.1016/j.cell.2012.05.044)
6. Karr JR et al. 2015 Summary of the DREAM8 parameter estimation challenge: toward parameter identification for whole-cell models. *PLoS Comput. Biol.* 11, e1004096. (doi:10.1371/journal.pcbi.1004096)
7. Lang M, Stelling J. 2016 Modular parameter identification of biomolecular networks. *SIAM J. Sci. Comput.* 38, B988–B1000. (doi:10.1137/15M103306X)
8. Babtie AC, Stumpf MPH. 2017 How to deal with complex biological networks: distinguishability issues and optimization-based solutions. *BMC Syst. Biol.* 5, 177. (doi:10.1186/s12918-015-0099-9)
9. Sonnhammer E, Sander C, Valencia A. 1995 A probabilistic model to infer protein secondary structures. *Bioinformatics* 11, 59–70. (doi:10.1093/bioinformatics/11.2.59)
10. Babtie AC, Kirk P, Stumpf MPH. 2014 Topological augmentation to infer hidden parameters in systems biology. *J. R. Soc. Interface* 11, 18507–18512. (doi:10.1073/pnas.1414026112)
11. Babtie AC, Kirk P, Stumpf MPH. 2014 Topological sensitivity analysis for systems biology. *Proc. Natl Acad. Sci. USA* 111, 18507–18512. (doi:10.1073/pnas.1414026112)
12. Leon M, Woods ML, Fedors A, Barnes CP. 2016 A computational method for the investigation of multistable systems and its application to genetic switches. *BMC Syst. Biol.* 10, 130–112. (doi:10.1186/s12918-016-0375-z)
13. Babtie AC, Kirk P, Stumpf MPH. 2014 Topological sensitivity analysis for systems biology. *Proc. Natl Acad. Sci. USA* 111, 18507–18512. (doi:10.1073/pnas.1414026112)
14. Leon M, Woods ML, Fedors A, Barnes CP. 2016 A computational method for the investigation of multistable systems and its application to genetic switches. *BMC Syst. Biol.* 10, 130–112. (doi:10.1186/s12918-016-0375-z)
15. Gersdorff J, Reddy NR, Lim WA. 2019 The design principles of biochemical timers: circuits that discriminate between transient and sustained stimulation. *Cell Syst.* 10, 100651. (doi:10.1016/j.cels.2019.07.008)
16. Scholes NS, Schnoor D, Isalan M, Stumpf MPH. 2019 A comprehensive network atlas reveals that tuning patterns are common but not
