Supplementary Material: Modeling Probability Forecasts via Information Diversity

Ville A. Satopää, Robin Pemantle, and Lyle H. Ungar

APPENDIX A: PROOFS AND DERIVATIONS

A.1 Proof of Proposition 3.3.

Denote the set of all coherent information structures with $\mathcal{Q}_N$. Consider $\mathbf{\Sigma}_{22} \in \mathcal{Q}_N$ and its associated Borel sets $\{B_i : i = 1, \ldots, N\}$. Given that $\mathbf{\Sigma}_{22}$ is coherent, its information can be represented in a diagram such as the one given by Figure 1 in the main manuscript. Keeping the diagram representation in mind, partition the unit interval S into 2^N disjoint parts $C_{\mathbf{v}} := \cap_{i \in \mathbf{v}} B_i \setminus \cup_{i \notin \mathbf{v}} B_i$, where $\mathbf{v} \subseteq \{1, \ldots, N\}$ denotes a subset of forecasters and each $C_{\mathbf{v}}$ represents information used only by the forecasters in $\mathbf{v}$. Given that $\sum_{\mathbf{v}} |C_{\mathbf{v}}| = 1$, it is possible to establish a linear function L from the probability simplex

$$\begin{aligned} \Delta_N &:= \text{conv}\{\boldsymbol{e}_{\mathbf{v}} : \mathbf{v} \subseteq \{1, \ldots, N\}\} \\ &= \left\{ \boldsymbol{z} \in \mathbb{R}^{2^N} : \boldsymbol{z} \geq \mathbf{0}, \mathbf{1}'\boldsymbol{z} = 1 \right\} \end{aligned}$$

to the space of coherent information structures $\mathcal{Q}_N$. In particular, the linear function $L : \boldsymbol{z} \in \Delta_N \rightarrow \mathbf{\Sigma}_{22} \in \mathcal{Q}_N$ is defined such that $\rho_{ij} = \sum_{\{i,j\} \subseteq \mathbf{v}} z_{\mathbf{v}}$ and $\delta_i = \sum_{i \in \mathbf{v}} z_{\mathbf{v}}$. Therefore

$L(\Delta_N) = \mathcal{Q}_N$. Furthermore, given that Δ_N is a convex polytope,

$$\begin{aligned} L(\Delta_N) &= \text{conv}\{L(\boldsymbol{e}_{\mathbf{v}}) : \mathbf{v} \subseteq \{1, \ldots, N\}\} \\ &= \text{conv}\left\{\boldsymbol{x}\boldsymbol{x}' : \boldsymbol{x} \in \{0,1\}^N\right\} \\ &= \text{COR}(N), \end{aligned} \tag{6}$$

which establishes $\text{COR}(N) = \mathcal{Q}_N$. Equality (6) follows from the basic properties of convex polytopes (see, e.g., McMullen and Shephard 1971, pp. 16). Each $\boldsymbol{\Sigma}_{22} \in \text{COR}(N)$ has $\frac{N(N+1)}{2} = \binom{n+1}{2}$ parameters and therefore exists in $\binom{n+1}{2}$ dimensions. □

A.2 Proof of Proposition 4.1

The proposition is proved by showing $\mathbb{E}(\mathbf{1}_A|\{X_{B_i}\}_{i=1}^N, X_{B'}) = \mathbb{E}(\mathbf{1}_A|X_{B'})$. First, append $X_{B'}$ to the multivariate Gaussian distribution (2) of the main manuscript:

$$\begin{pmatrix} X_S \\ X_{B'} \\ X_{B_1} \\ X_{B_2} \\ \vdots \\ X_{B_N} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \Omega_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix} = \left(\begin{array}{c|cccccc} 1 & \delta' & \delta_1 & \delta_2 & \ldots & \delta_N \\ \hline \delta' & \delta' & \delta_1 & \delta_2 & \ldots & \delta_N \\ \delta_1 & \delta_1 & \delta_1 & \rho_{1,2} & \ldots & \rho_{1,N} \\ \delta_2 & \delta_2 & \rho_{2,1} & \delta_2 & \ldots & \rho_{2,N} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta_N & \delta_N & \rho_{N,1} & \rho_{N,2} & \ldots & \delta_N \end{array}\right)\right).$$

Denote $\mathbf{X}_\Omega = (X_{B'}, X_{B_1}, \ldots, X_{B_N})'$. If $\mathbf{e}_1$ is the first standard basis vector of length $N+1$ and the above multivariate Gaussian distribution is non-degenerate, then $\boldsymbol{\Omega}_{21} = \mathbf{e}_1'\boldsymbol{\Omega}_{22} \Leftrightarrow \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{22}^{-1} = \mathbf{e}_1'$. This identity together with the well-known results of the conditional Gaussian

distributions (see, e.g., Ravishanker and Dey 2001, Result 5.2.10) give

$$\begin{aligned}\mathbb{E}(\mathbf{1}_A|\{X_{B_i}\}_{i=1}^N, X_{B'}) &= \Phi\left(\frac{\mathbf{\Omega}_{12}\mathbf{\Omega}_{21}^{-1}\mathbf{X}_\Omega}{\sqrt{1-\mathbf{\Omega}_{12}\mathbf{\Omega}_{21}^{-1}\mathbf{\Omega}_{21}}}\right)\\ &= \Phi\left(\frac{\mathbf{e}_1'\mathbf{X}_\Omega}{\sqrt{1-\mathbf{e}_1'\mathbf{\Omega}_{21}}}\right)\\ &= \Phi\left(\frac{X_{B'}}{\sqrt{1-\delta'}}\right)\\ &= \mathbb{E}(\mathbf{1}_A|X_{B'})\end{aligned}$$

□

A.3 Proof of Proposition 4.2.

Given that

$$\begin{aligned}P' &\sim \mathcal{N}\left(0, \sigma_1^2 := \frac{\delta'}{1-\delta'}\right)\\ \frac{1}{N}\sum_{i=1}^N P_i &\sim \mathcal{N}\left(0, \sigma_2^2 := \frac{1}{N^2}\left\{\sum_{i=1}^N \frac{\delta_i}{1-\delta_i} + 2\sum_{i,j:i<j}\frac{\rho_{ij}}{\sqrt{(1-\delta_j)(1-\delta_i)}}\right\}\right),\end{aligned}$$

the amount of extremizing α is a ratio of two correlated Gaussian random variables. The Pearson product-moment correlation coefficient for them is

$$\kappa = \frac{\sum_{i=1}^N \frac{\delta_i}{\sqrt{1-\delta_i}}}{\sqrt{\delta'\left\{\sum_{i=1}^N \frac{\delta_i}{1-\delta_i} + 2\sum_{i,j:i<j}\frac{\rho_{ij}}{\sqrt{(1-\delta_j)(1-\delta_i)}}\right\}}}$$

It follows that α has a Cauchy distribution as long as $\sigma_1 \neq 1$, $\sigma_2 \neq 1$, or $\kappa \pm 1$ (see, e.g., Cedilnik et al. 2004). These conditions are very mild under the Gaussian model. For instance,

if no forecaster knows as much as the oracle, the conditions are satisfied. Consequently, the probability density function of α is

$$f(\alpha|x_0,\gamma) = \frac{1}{\pi}\frac{\gamma}{(\alpha - x_0)^2 + \gamma^2},$$

where $x_0 = \kappa\sigma_1/\sigma_2$ and $\gamma = \sqrt{1-\kappa^2}\sigma_1/\sigma_2$. The parameter x_0 represents the location (the median and mode) and γ specifies the scale (half the interquartile range) of the Cauchy distribution. The location parameter simplifies to

$$x_0 = \kappa\frac{\sigma_1}{\sigma_2} = \frac{N\sum_{i=1}^{N}\frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta')}}}{\sum_{i=1}^{N}\frac{\delta_i}{1-\delta_i} + 2\sum_{i,j:i<j}\frac{\rho_{ij}}{\sqrt{(1-\delta_j)(1-\delta_i)}}}$$

Given that all the remaining terms are positive, the location parameter x_0 is also positive. Compare the N terms with a given subindex i in the numerator with the corresponding terms in the denominator. From $\delta' \geq \delta_i \geq \rho_{ij}$, it follows that

$$\frac{\delta_i}{1-\delta_i} = \frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta_i)}} \leq \frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta')}} \tag{7}$$

$$\frac{\rho_{ij}}{\sqrt{(1-\delta_j)(1-\delta_i)}} \leq \frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta')}} \tag{8}$$

Therefore

$$N\sum_{i=1}^{N}\frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta')}} \geq \sum_{i=1}^{N}\frac{\delta_i}{1-\delta_i} + 2\sum_{i,j:i<j}\frac{\rho_{ij}}{\sqrt{(1-\delta_j)(1-\delta_i)}},$$

which gives that $x_0 \geq 1$. Given that the Cauchy distribution is symmetric around x_0, it must be the case that $\mathbb{P}(\alpha > 1|\mathbf{\Sigma}_{22},\delta') \geq 1/2$. Based on (7) and (8), the location $x_0 = 1$ only when all the forecasters know the same information, i.e., when $\delta_i = \delta_j$ for all $i \neq j$. Under

this particular setting, the amount of extremizing α is non-random and always equal to one. Any deviation from this particular information structure makes α random, $x_0 > 1$, and hence $\mathbb{P}(\alpha > 1|\boldsymbol{\Sigma}_{22}, \delta') > 1/2$. □

A.4 Derivation of Equation 4

Clearly, any $\delta \in [0, 1]$ is plausible. Conditional on such δ, however, the overlap parameter λ must be within a subinterval of $[0, 1]$. The upper bound of this subinterval is always one because the forecasters may use the same information under any δ and N. To derive the lower bound, note that information overlap is unavoidable when $\delta > 1/N$, and that minimum overlap occurs when all information is used either by everyone or by a single forecaster. In other words, if $\delta > 1/N$ and $B_i \cap B_j = B$ with $|B| = \lambda\delta$ for all $i \neq j$, the value of λ is minimized when $\lambda\delta + N(\delta - \delta\lambda) = 1$. Therefore the lower bound for λ is $\max\{(N - \delta^{-1})/(N - 1), 0\}$, and $\boldsymbol{\Sigma}_{22}$ is coherent if and only if $\delta \in [0, 1]$ and $\lambda|\delta \in [\max\{(N - \delta^{-1})/(N - 1), 0\}, 1]$.

A.5 Proof of Proposition 5.1.

(i) This follows from direct computation:

$$\begin{aligned}\alpha &= \left(\frac{\frac{1}{(N-1)\lambda+1}\sum_{i=1}^{N} X_{B_i}}{\sqrt{1-\frac{N\delta}{(N-1)\lambda+1}}}\right) \Bigg/ \left(\frac{1}{N}\sum_{i=1}^{N}\frac{X_{B_i}}{\sqrt{1-\delta}}\right) \\ &= \frac{\frac{N\sqrt{1-\delta}}{(N-1)\lambda+1}}{\sqrt{1-\frac{N\delta}{(N-1)\lambda+1}}}, \end{aligned} \tag{9}$$

which simplifies to the given expression after substituting in γ. Given that this quantity does not depend on any X_{B_i}, it is non-random.

(ii) For a given δ, the amount of extremizing α is minimized when $(N-1)\lambda+1$ is maximized.

This happens as $\lambda \uparrow 1$. Plugging this into (9) gives

$$\alpha = \frac{\frac{N\sqrt{1-\delta}}{(N-1)\lambda+1}}{\sqrt{1-\frac{N\delta}{(N-1)\lambda+1}}} \downarrow \frac{\sqrt{1-\delta}}{\sqrt{1-\delta}} = 1$$

(iii) Assume without loss of generality that $\bar{P} > 0$. If $\max\{p_1, p_2, \ldots, p_N\} < 1$, then setting $\delta = 1/N$ and $\lambda = 0$ gives an aggregate probability $p'' = 1$ that is outside the convex hull of the individual probabilities. □

APPENDIX B: PARAMETER ESTIMATION UNDER SYMMETRIC INFORMATION

This section describes how the maximum likelihood estimates of δ and λ can be found accurately and efficiently. Denote a $N \times N$ matrix of ones with $\mathbf{J}_N$. A matrix $\mathbf{\Sigma}$ is called compound symmetric if it can be expressed in the form $\mathbf{\Sigma} = \mathbf{I}_N A + \mathbf{J}_N B$ for some constants A and B. The inverse matrix (if it exists) and any scalar multiple of a compound symmetric matrix $\mathbf{\Sigma}$ are also compound symmetric (Dobbin and Simon, 2005). More specifically, for some constant c,

$$\begin{aligned} c\mathbf{\Sigma} &= \mathbf{I}_N(cA) + \mathbf{J}_N(cB) \\ \mathbf{\Sigma}^{-1} &= \mathbf{I}_N \frac{1}{A} - \mathbf{J}_N \frac{B}{A(A+NB)} \end{aligned} \tag{10}$$

Define

$$\begin{aligned} \mathbf{\Sigma}_{22} &:= \mathrm{Cov}\,(\boldsymbol{X}) = \mathbf{I}_N A_X + \mathbf{J}_N B_X \\ \mathbf{\Sigma}_P &:= \mathrm{Cov}\,(\boldsymbol{P}) = \mathbf{\Sigma}_{22}/(1-\delta) = \mathbf{I}_N A_P + \mathbf{J}_N B_P \\ \mathbf{\Omega} &:= \mathbf{\Sigma}_P^{-1} = \mathbf{I}_N A_\Omega + \mathbf{J}_N B_\Omega \end{aligned} \tag{11}$$

To set up the optimization problem, observe that the Jacobian for the map $\boldsymbol{P} \rightarrow \Phi(\boldsymbol{P}) = (\Phi(P_1), \Phi(P_2), \ldots, \Phi(P_N))'$ is $J(\boldsymbol{P}) = (2\pi)^{-N/2} \exp(-\boldsymbol{P}'\boldsymbol{P}/2)$. If $h(\boldsymbol{P})$ denotes the multivariate Gaussian density of $\boldsymbol{P} \sim \mathcal{N}_N(\mathbf{0}, \boldsymbol{\Sigma}_P)$, the density for $\boldsymbol{p} = (p_1, p_2, \ldots, p_N)'$ is

$$f(\boldsymbol{p}|\delta, \lambda) \;\; = \;\; h(\boldsymbol{P})J(\boldsymbol{P})^{-1} \propto |\boldsymbol{\Sigma}_P|^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{P}'\boldsymbol{\Sigma}_P^{-1}\boldsymbol{P}\right],$$

where $\boldsymbol{P} = \Phi^{-1}(\boldsymbol{p})$. Let $\mathbf{S}_P = \boldsymbol{P}\boldsymbol{P}'$ be the (rank one) sample covariance matrix of $\boldsymbol{P}$. The log-likelihood then reduces to

$$\log f(\boldsymbol{p}|\delta, \lambda) \propto -\log\det\boldsymbol{\Sigma}_P - \operatorname{tr}\left(\mathbf{S}_P^{-1}\boldsymbol{\Sigma}_P\right)$$

This log-likelihood is not concave in $\boldsymbol{\Sigma}_P$. It is, however, a concave function of $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_P^{-1}$. Making this change of variables gives us the following optimization problem:

$$\begin{aligned} &\text{minimize} \;\; -\log\det\boldsymbol{\Omega} + \operatorname{tr}(\mathbf{S}_P\boldsymbol{\Omega}) \\ &\text{subject to } \delta \in [0,1] \\ &\qquad\qquad \lambda \in \left[\max\left\{\frac{N-\delta^{-1}}{N-1}, 0\right\}, 1\right), \end{aligned} \tag{12}$$

where the open upper bound on λ ensures a non-singular information structure $\boldsymbol{\Sigma}_{22}$. Unfortunately, the feasible region is not convex (see, e.g., Figure 3 in the main manuscript) but can be made convex by re-expressing the problem as follows: First, let $\rho = \delta\lambda$ denote the amount of information known by a forecaster; that is, let $A_X = (\delta - \rho)$ and $B_X = \rho$. Solving the problem in terms of δ and ρ is equivalent to minimizing the original objective (12) but subject to $0 \leq \rho \leq \delta$ and $0 \leq \rho(N-1) - N\delta + 1$. Given that this region is an intersection of four half-spaces, it is convex. Furthermore, it can be translated into the corresponding feasible and

convex set of (A_Ω, B_Ω) via the following steps:

$$
\begin{aligned}
& \boldsymbol{\Sigma}_{22} \in \{\boldsymbol{\Sigma}_{22} : 0 \leq \rho \leq \delta, 0 \leq \rho(N-1) - N\delta + 1\} \\
\Leftrightarrow \quad & \boldsymbol{\Sigma}_{22} \in \{\boldsymbol{\Sigma}_{22} : 0 \leq B_X, 0 \leq A_X, 0 \leq 1 - B_X + NA_X, \} \\
\Leftrightarrow \quad & \boldsymbol{\Sigma}_P \in \{\boldsymbol{\Sigma}_P : 0 \leq A_P \leq 1/(N-1), 0 \leq B_P\} \\
\Leftrightarrow \quad & \boldsymbol{\Omega} \in \{\boldsymbol{\Omega} : 0 \leq A_\Omega - N + 1, 0 \leq A_\Omega + B_\Omega N, 0 \leq -B_\Omega\}
\end{aligned}
$$

According to Rao (2009), $\log\det(\boldsymbol{\Omega}) = N\log A_\Omega + \log\left(1 + NB_\Omega/A_\Omega\right)$. Plugging this and the feasible region of (A_Ω, B_Ω) into the original problem (12) gives an equivalent but convex optimization problem:

$$
\begin{aligned}
\text{minimize } & -N\log A_\Omega - \log\left(1 + \frac{NB_\Omega}{A_\Omega}\right) + A_\Omega \operatorname{tr}(\mathbf{S}_P) + B_\Omega \operatorname{tr}(\mathbf{S}_P\mathbf{J}_N) \\
\text{subject to } & 0 \leq A_\Omega - N + 1 \\
& 0 \leq A_\Omega + B_\Omega N \\
& 0 \leq -B_\Omega
\end{aligned}
$$

The first term of this objective is both convex and non-decreasing. The second term is a composition of the same convex, non-decreasing function with a function that is concave over the feasible region. Such a composition is always convex. The last two terms are affine and hence also convex. Therefore, given that the objective is a sum of four convex functions, it is convex, and globally optimal values of (A_Ω, B_Ω) can be found very efficiently with interior point algorithms such as the barrier method. There are many open software packages that implement generic versions of these methods. For instance, our implementation uses the standard R function `constrOptim` to solve the optimization problem. Denote optimal values with (A_Ω^*, B_Ω^*).

They can be traced back to (δ, λ) via (10) and (11). The final map simplifies to

$$\delta^* = \frac{B^*_\Omega(N-1) + A^*_\Omega}{A^*_\Omega(1+A^*_\Omega) + B^*_\Omega(N-1+NA^*_\Omega)} \qquad \text{and} \qquad \lambda^* = -\frac{B^*_\Omega}{B^*_\Omega(N-1) + A^*_\Omega}$$

REFERENCES

Cedilnik, A., Kosmelj, K., and Blejec, A. (2004). The distribution of the ratio of jointly normal variables. *Metodoloski Zvezki*, 1(1):99–108.

Dobbin, K. and Simon, R. (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, 6(1):27–38.

McMullen, P. and Shephard, G. C. (1971). *Convex Polytopes and the Upper Bound Conjecture*, volume 3. Cambridge University Press, Cambridge, U.K.

Rao, C. R. (2009). *Linear Statistical Inference and Its Applications*, volume 22 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, New York, New York.

Ravishanker, N. and Dey, D. K. (2001). *A first course in linear model theory*. CRC Press.