Data Fusion with Confidence Curves: The II-CC-FF Paradigm

Nils Lid Hjort (with Céline Cunen)

Department of Mathematics, University of Oslo

BFF4, Harvard, May 2017

The problem: Combining information

Suppose ψ is a parameter of interest, with data $y_1, \ldots, y_k$ from sources $1, \ldots, k$ carrying information about ψ. How to combine these pieces of information?

Standard (and simple) example: $y_j \sim \mathrm{N}(\psi, \sigma_j^2)$ are independent, with known or well estimated σ_j. Then

$$\widehat{\psi} = \frac{\sum_{j=1}^k y_j/\sigma_j^2}{\sum_{j=1}^k 1/\sigma_j^2} \sim \mathrm{N}\Big(\psi, \frac{1}{\sum_{j=1}^k 1/\sigma_j^2}\Big).$$

Often additional variability among the ψ_j. Would e.g. be interested in assessing both parameters of $\psi \sim \mathrm{N}(\psi_0, \tau^2)$.

We need extended methods and partly new paradigms for handling cases with very different types of information.

Plan

General problem formulation:

Data y_j source j carry information about ψ_j. Wish to assess overall aspects of these ψ_j, perhaps for inference concerning some $\phi = \phi(\psi_1, \ldots, \psi_k)$.

- A Confidence distributions & confidence curves
- B Previous CD combination methods (Singh, Strawderman, Xie, Liu, Liu, others)
- C A different II-CC-FF paradigm, via steps Independent Inspection, Confidence Conversion, Focused Fusion, and confidence-to-likelihood operations
- D1 Example 1: Meta-analysis for 2×2 tables
- D2 Example 2: Effective population size for cod
- D3 Example 3: Farmed salmons escaping to Norwegian rivers
- D4 Example 4: Abundance assessment for Humpback Whales
- E Concluding remarks

A: Confidence distributions and confidence curves

For a parameter ψ, suppose data y give rise to confidence intervals, say $[\psi_{0.05}, \psi_{0.95}]$ at level 0.90, but also for other levels. These are converted into a full distribution of confidence, with

$$[\psi_{0.05}, \psi_{0.95}] = [C^{-1}(0.05, y_{\rm obs}), C^{-1}(0.95, y_{\rm obs})],$$

etc. Here $C(\psi, y)$ is a cdf in ψ, for each y, and

$$C(\psi_0, Y) \sim {\rm unif} \quad \text{at true value } \psi_0.$$

Very useful, also qua graphical summary: the confidence curve ${\rm cc}(\psi, y_{\rm obs})$ with

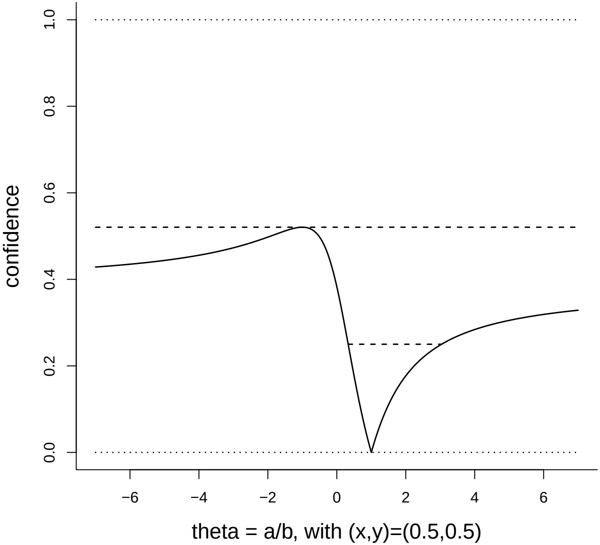
$${\rm Pr}_\psi\{{\rm cc}(\psi, Y) \le \alpha\} = \alpha \quad \text{for all } \alpha.$$

From CD to cc: ${\rm cc}(\psi) = |1 - 2\,C(\psi, y_{\rm obs})|$, with ${\rm cc}(\psi) = 0.90$ giving the two roots $\psi_{0.05}, \psi_{0.95}$, etc. The cc is more fundamental than the CD.

An extensive theory is available for CD and cc, cf. Confidence, Likelihood, Probability, Schweder and Hjort (CUP, 2016).

Cox (yesterday): $x \sim \mathrm{N}(a, 1)$, $y \sim \mathrm{N}(b, 1)$, observe $(0.50, 0.50)$: the canonical confidence curve for $\theta = a/b$ is

$$\mathrm{cc}(\theta) = \Gamma_1\Big(\frac{(x-\theta y)^2}{1+\theta^2}\Big).$$



B: Liu, Liu, Singh, Strawderman, Xie et al. methods

Data y_j give rise to a CD $C_j(\psi, y_j)$ for ψ. Under true value, $C_j(\psi, Y_j) \sim \mathrm{unif}$. Hence $\Phi^{-1}(C_j(\psi, Y_j)) \sim \mathrm{N}(0,1)$, and

$$\bar{C}(\psi) = \Phi\Big(\sum_{j=1}^{k} w_j \Phi^{-1}(C_j(\psi, Y_j))\Big)$$

is a combined CD, if the weights w_j are nonrandom and $\sum_{j=1}^k w_j^2 = 1$.

This is a versatile and broadly applicable method, but with some drawbacks: (a) trouble when estimated weights $\widehat{w}_j$ are used; (b) lack of full efficiency. In various cases, there are better CD combination methods, with higher confidence power.

Better (in various cases): sticking to likelihoods and sufficiency.

CD (or cc) combination via confidence likelihoods

Combining information, for inference about focus parameter $\phi = \phi(\psi_1, \ldots, \psi_k)$: General II-CC-FF paradigm for combination of information sources:

II: Independent Inspection: From data source y_j to estimate and intervals, yielding a cc:

$$y_j \Longrightarrow \mathrm{cc}_j(\psi_j).$$

CC: Confidence Conversion: From the confidence curve to a confidence log-likelihood,

$$\mathrm{cc}_j(\psi_j) \Longrightarrow \ell_{c,j}(\psi_j).$$

FF: Focused Fusion: Use the combined confidence log-likelihood $\ell_c = \sum_{j=1}^k \ell_{c,j}(\psi_j)$ to construct a cc for the given focus $\phi = \phi(\psi_1, \ldots, \psi_k)$, perhaps via profiling, median-Bartletting, etc.:

$$\ell_c(\psi_1, \ldots, \psi_k) \Longrightarrow \bar{\mathrm{cc}}_{\mathrm{fusion}}(\phi).$$

FF is also the (focused) Summary of Summaries operation.

Carrying out steps II, CC, FF can be hard work, depending on circumstances. The CC step is sometimes the hardest (conversion of CD or cc to log-likelihood). The simplest method is normal conversion,

$$\ell_{c,j}(\psi_j) = -\tfrac{1}{2}\,\Gamma_1^{-1}(\mathrm{cc}_j(\psi_j)) = -\tfrac{1}{2}\{\Phi^{-1}(C_j(\psi_j))\}^2,$$

but more elaborate methods may typically be called for.

Sometimes step II needs to be based on summaries from other work (e.g. from point estimate and a .95 interval to approximate CD).

With raw data and sufficient time for careful modelling, steps II and CC may lead to $\ell_{c,j}(\psi_j)$ directly. Even then having individual CDs for the ψ_j is informative and useful.

Illustration 1: Classic meta-analysis.

II: Independent Inspection: Statistical work with data source y_j leads to $\widehat\psi_j \sim \mathrm{N}(\psi_j, \sigma_j^2)$; $C_j(\psi_j) = \Phi((\psi_j - \widehat\psi_j)/\sigma_j)$.

CC: Confidence Conversion: From $C_j(\psi_j)$ to $\ell_{c,j}(\psi_j) = -\frac{1}{2}(\psi_j - \widehat\psi_j)^2/\sigma_j^2$.

FF: Focused Fusion: With a common mean parameter across studies: Summing $\ell_{c,j}(\psi_j)$ leads to classic answer

$$\widehat\psi = \frac{\sum_{j=1}^k \widehat\psi_j/\sigma_j^2}{\sum_{j=1}^k 1/\sigma_j^2} \sim \mathrm{N}\Big(\psi, \Big(\sum_{j=1}^k 1/\sigma_j^2\Big)^{-1}\Big).$$

With ψ_j varying as $\mathrm{N}(\psi_0, \tau^2)$: then $\widehat\psi_j \sim \mathrm{N}(\psi_0, \tau^2 + \sigma_j^2)$. CD for τ:

$$C(\tau) = \mathrm{Pr}_\tau\{Q_k(\tau) \ge Q_{k,\mathrm{obs}}(\tau)\} = 1 - \Gamma_{k-1}(Q_{k,\mathrm{obs}}(\tau)),$$

with $Q_k(\tau) = \sum_{j=1}^k \{\widehat\psi_j - \bar\psi(\tau)\}^2/(\tau^2 + \sigma_j^2)$. There is a positive confidence probability for $\tau = 0$. CD for ψ_0: based on t-bootstrapping and

$$t = \{\bar\psi(\widehat\tau) - \psi\}/\kappa(\widehat\tau).$$

Illustration 2: Let $Y_j \sim \mathrm{Gamma}(a_j, \theta)$, with known shape a_j.

II: Independent Inspection: Optimal CD for θ based in Y_j is $C_j(\theta) = G(\theta y_j, a_j, 1)$.

CC: Confidence Conversion: From $C_j(\theta)$ to $\ell_{c,j}(\psi_j) = -\theta y_j + a_j \log\theta$.

FF: Focused Fusion: Summing confidence log-likelihoods, $\bar{C}_{\mathrm{fusion}}(\theta) = G(\theta \sum_{j=1}^{k} y_j, \sum_{j=1}^{k} a_j, 1)$. This is the optimal CD for θ, and has higher CD performance than the Singh, Strawderman, Xie type

$$\widetilde{C}(\theta) = \Phi\Big(\sum_{j=1}^{k} w_j \Phi^{-1}(C_j(\theta)\Big),$$

even for the optimally selected w_j.

Crucially, the II-CC-FF strategy is very general and can be used with very different data sources (e.g. hard and soft and big and small data). The potential of the II-CC-FF paradigm lies in its use for much more challenging applications (where each of II, CC, FF might be hard).

D1: Meta-analysis for 2×2 tables with incomplete information

Death rates for two groups of acute myocardial infarction patients, with the 2nd group using Lidocaine:

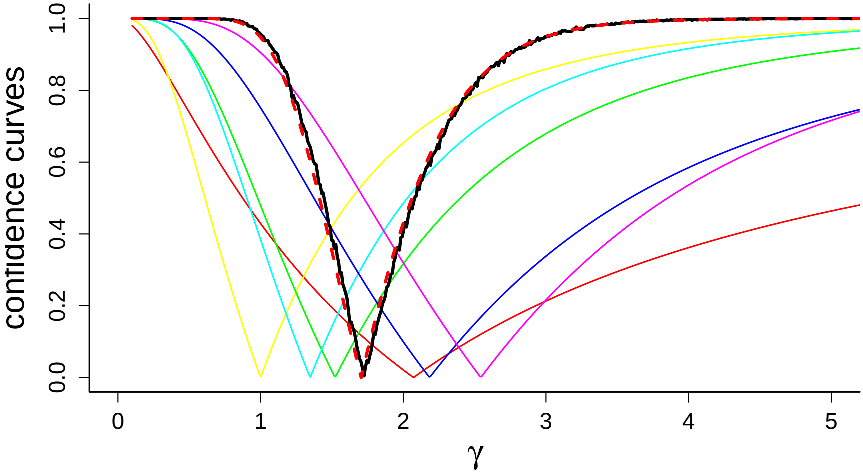
$$y_{i,0} \sim \mathrm{Pois}(e_{i,0}\lambda_{i,0}) \quad \text{and} \quad y_{i,1} \sim \mathrm{Bin}(e_{i,1}\lambda_{i,1}),$$

with $e_{i,0}$ and $e_{i,1}$ exposure numbers (proportional to sample sizes) and

$$\lambda_{i,1} = \gamma\lambda_{i,0} \quad \text{for } i = 1, \ldots, 6.$$

m_1	m_0	y_1	y_0
39	43	2	1
44	44	4	4
107	110	6	4
103	100	7	5
110	106	7	3
154	146	11	4

I (i) produce optimal $\mathrm{cc}_j(\gamma)$ for each of the six studies; (ii) combine these using II-CC-FF; (iii) also compute gold standard $\mathrm{cc}(\gamma)$. Results of (ii) and (iii) are amazingly close.



D2: Effective population size ratio for cod

A certain population of cod is studied. Of interest is both actual population size N and effective population size N_e (the size of a hypothetical stable population, with the same genetic variability as the full population, and where each individual has a binomially distributed number of reproducing offspring). The biological focus parameter in this study is $\phi = N_e/N$.

Steps II-CC for N: A CD for N, with confidence log-likelihood: A certain analysis leads to confidence log-likelihood

$$\ell_c(N) = -\tfrac{1}{2}(N - 1847)^2/534^2.$$

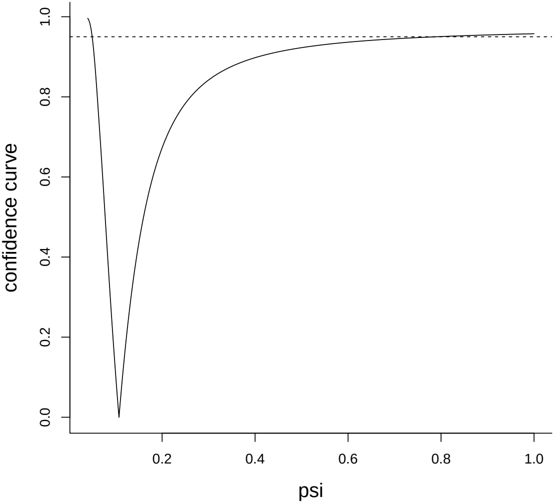
Steps II-CC for N_e: A CD for N_e, with confidence log-likelihood: This is harder, via genetic analyses, etc., but yields confidence log-likelihood

$$\ell_{c,e}(N_e) = -\tfrac{1}{2}(N_e^b - 198^b)/s^2$$

for certain estimated transformation parameters (b, s).

Step FF for the ratio: A CD for $\phi = N_e/N$. This is achieved via log-likelihood profiling and median-Bartletting,

$$\ell_{\rm prof}(\phi) = \max\{\ell_c(N) + \ell_{c,e}(N_e)\colon N_e/N = \phi\}.$$



D3: Wild salmon vs. farmed salmon in Norwegian rivers

Substantial amounts of farmed salmon escape and are found in 'the wild'. One wishes to estimate

$$p = \Pr(A),$$

the proportion of escapees in a river. Cathcing m salmon and finding y of these are from the farmed population gives information on p' not p:

$$\begin{aligned} p' &= \Pr(\text{farmed} \,|\, \text{caught}) \\ &= \frac{p \Pr(\text{caught} \,|\, \text{farmed})}{p \Pr(\text{caught} \,|\, \text{farmed}) + (1-p) \Pr(\text{caught} \,|\, \text{wild})} \\ &= \frac{p\,\rho}{p\,\rho + 1 - p} = h(p, \rho), \end{aligned}$$

with

$$\rho = \frac{\Pr(\text{caught} \,|\, \text{farmed})}{\Pr(\text{caught} \,|\, \text{wild})}.$$

How to turn information on $p'_j = h(p_j, \rho)$ into information on p_j?

So we have $y_j \sim \mathrm{Bin}(m_j, p'_j)$ for rivers $j = 1, \ldots, k$ with

$$p'_j = \frac{p_j \rho}{p_j \rho + 1 - p_j} = h(p_j, \rho).$$

The fisheries scientists have *some* information on ρ, which we translate to a CD, then to $\ell_0(\rho)$. With binomial log-likelihoods for p'_j, we use

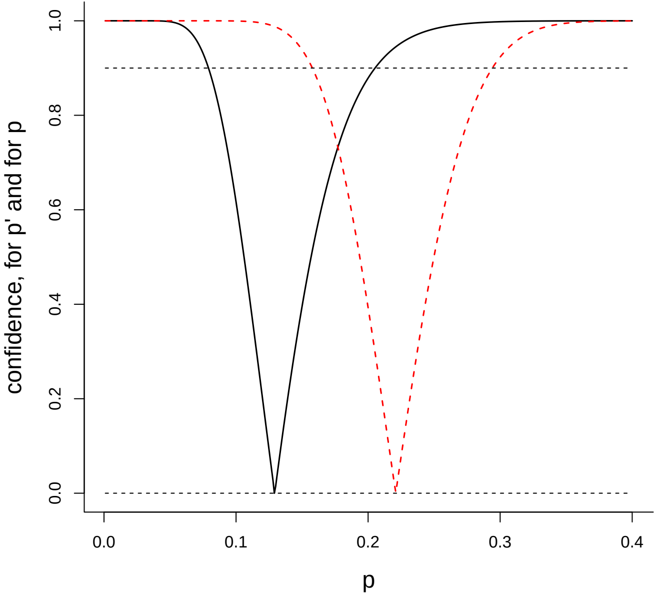
$$\ell(p_1, \ldots, p_k, \rho) = \sum_{j=1}^{k} \ell_j(h(p_j, \rho)) + \ell_0(\rho)$$

and base further inference on

$$\ell_{\rm prof}(p_1, \ldots, p_k) = \max\Big\{\sum_{j=1}^{k} \ell_j(h(p_j, \rho)) + \ell_0(\rho) \colon \text{all } \rho\Big\}.$$

The CD for ρ is close to that of a posterior $\mathrm{Gamma}(22, 11)$ with mean 2.

With $y = 22$ from $\mathrm{Bin}(100, p')$, $p' = h(p, \rho)$, and via separate $\mathrm{cc}_0(\rho)$: estimate shifted from 0.22 to 0.13, along with confidence curve.



Note: $\exists$ many other setups where the real need is inference for $p = \Pr(A)$, but first-line data instead pertain to

$$p' = \Pr(A') = \text{some } h(p, \rho).$$

Not hard: estimating $p' = \Pr(A')$ with
A': person says or thinks he or she will vote DT in three weeks

To get to
A: person will actually vote DT once election is there
we would need separate information on

$$\rho = \Pr(\text{person will vote DT} \,|\, \text{person says he will vote HC}).$$

With such information, could do II-CC-FF to go from $\mathrm{cc}_j(p'_j)$ (around 0.33?) to $\mathrm{cc}_j(p_j)$ (around 0.52?).

D4: Humpback Whales

Abundance assessment of a humpback population, from 1995 and 2001, summarised as 2.5%, 50%, 97.5% confidence quantiles – from two separate studies, with very different types of data and very different statistical methods:

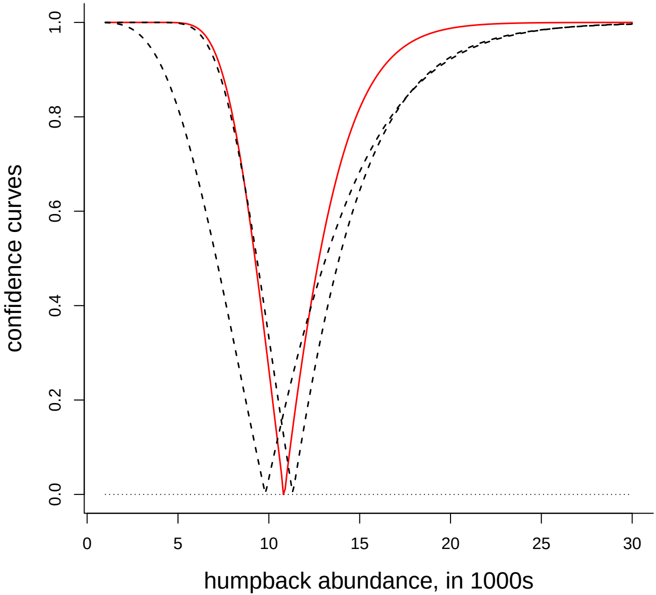
	2.5%	50%	97.5%
1995	3439	9810	21457
2001	6651	11319	21214

Note that intervals are skewed to the right.

Challenge: combining information, finding estimates and (approximate) confidence intervals (and a full confidence curve) for the population size N, assuming this to have been approximately constant.

Method: $C_j(N) = \Phi((h_j(N) - h_j(\widehat{N}_j))/s_j)$ with power transformation $h_j(N) = (1/a)(N^a - 1)$ $\Longrightarrow$ fitting $\Longrightarrow$ transforming to log-likelihoods $\Longrightarrow$ adding $\Longrightarrow$ converting.

This yields $\mathrm{cc}(N)$ for 1995 and 2001, and for the combined information: point estimate 10487, 95% interval $[6692, 17156]$.



E: Concluding remarks (and further questions)

a. Can handle parametric + nonparametric in the II-CC-FF scheme, as long as we have $\mathrm{cc}_j(\psi_j)$.

b. Can allow Bayesian components too:

$$\ell_c = \sum_{j=1}^{k} \ell_{c,j}(\psi_j) + \ell_0(\rho)$$

could have log-prior, og log-posterior, contributions. A log-prior or log-posterior for ρ can be taken on board without the full Bayesian job (of having a joint prior for all parameters of all models).

c. If we have the raw data, and have the time and resources to do all the full analyses ourselves, then we would find the $C_j(\psi_j)$ in Step II = Independent Inspection. In real world we would often only be able to find a point estimate and a 95% interval for the ψ_j. We may still squeeze an approximate CD out of this.

d. Step CC = Confidence Conversion is often tricky. There is no one-to-one correspondence between log-likelihoods and CDs. Data protocol matters. See CLP (2016).

e. Step FF = Focused Fusion may be accomplished by profiling the combined confidence log-likelihood, followed by fine-tuning (Bartletting, median correction, abc bootstrapping).

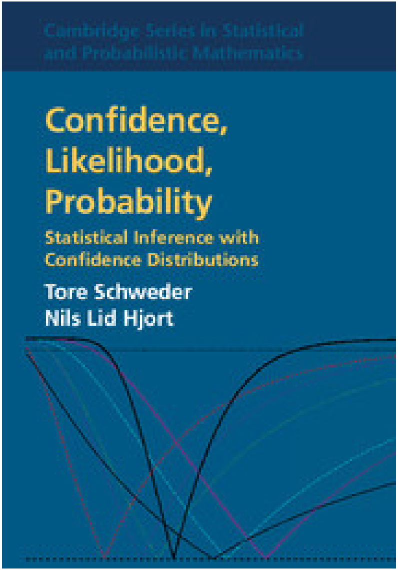
f. Other 'harder applications' of the II-CC-FF scheme are under way (inside the FocuStat research programme 2014–2018) – involving hard and soft data, as well as with big and small data.

– Who wins the 2018 Football World Cup? Combining FIFA ranking numbers with expert opinions, 1 day before each match. System will be in place, with day-to-day updating, June-July 2018.

– Evolutionary diversification rates for mammals over the past 40 million years: fossil records + phylogeny.

– Air pollution data for European cities, aiming at CDs for $\mathrm{Pr}(\text{tomorrow will be above threshold})$.

CLP, 2016:


Cambridge Series in Statistical
and Probabilistic Mathematics
Confidence,
Likelihood,
Probability
Statistical Inference with
Confidence Distributions
Tore Schweder
Nils Lid Hjort