IMS Collections
Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman
Vol. 8 (2012) 200–214
© Institute of Mathematical Statistics, 2012
DOI: 10.1214/11-IMSCOLL814

CD posterior – combining prior and data through confidence distributions

Kesar Singh and Minge Xie[*]

Rutgers, the State University of New Jersey

Abstract: This article proposes an alternative approach to incorporate information from observed data with its corresponding prior information using a recipe developed for combining confidence distributions. The outcome function is called a *CD posterior*, an alternative to Bayes posterior, which is shown here to have the same coverage property as the Bayes posterior. This approach to incorporating a prior distribution has a great advantage that it does not require any prior on nuisance parameters. It also can ease the computational burden which a typical Bayesian analysis endures. An error bound is established on the CD-posterior when there is an error in prior specification.

Contents

1. Introduction

The basic logic of objective Bayesian analysis is to bring together prior knowledge on a population parameter and the knowledge acquired on it out of a quantitative study. The prior knowledge, represented in the form of a probability distribution, which the parameter is assumed to follow, is supposedly an aggregate of past experiences, experts' training, expertise and even subjective opinions. After sample data are observed, a model is built which specifies the likelihood function of data given the values of the parameters. The prior and likelihood functions are incorporated together using the Bayes formula, and the resulting conditional distribution of the parameter given the data is a Bayes posterior. This article proposes an alternative way to bring together prior distribution and the data, introducing the concept of CD-posterior, where CD refers to the concept of confidence distribution.

501 Hill Center, Busch Campus, Department of Statistics and Biostatistics, Rutgers, The State University of New Jersey, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019

[*] We dedicate this article to our colleague Bill Strawderman. This research is supported in part by grants from the NSF, NSA and DHS.

AMS 2000 subject classifications: Primary 62A01, 62F03, 62F12, 62F15, 62F40, 62G05, 62G10, 62G20

Keywords and phrases: posterior distribution, confidence distribution, frequentist coverage

Suppose we are interested in making an inference on a given real valued parameter θ, which is allowed to be just a functional of the underlying population. Denote the sample data by $\mathbf{X}$ and the nuisance parameters by ϕ. Let $Pos(\theta)$ be the marginal posterior cumulative distribution function of θ, given observed data $\mathbf{X}$. Let $Pos^{-1}(t)$ be its t^{th} quantile that solves the equation $Pos(\theta) = t$. The credible intervals $[Pos^{-1}(t_1), Pos^{-1}(t_2)]$ has the following coverage probability,

$$P_{\theta|\mathbf{X}}\{Pos^{-1}(t_1) \le \theta \le Pos^{-1}(t_2)\} = t_2 - t_1 \tag{1.1}$$

for some $0 < t_1 < t_2 < 1$. Here, $P_{\theta|\mathbf{X}}$ refers to the conditional probability measure of θ given $\mathbf{X}$. Consequently, by averaging,

$$\begin{aligned} &P_{(\theta,\mathbf{X})}\{Pos^{-1}(t_1) \le \theta \le Pos^{-1}(t_2)\} \\ &= E_{\mathbf{X}}[P_{\theta|\mathbf{X}}\{Pos^{-1}(t_1) \le \theta \le Pos^{-1}(t_2)\}] \\ &= t_2 - t_1. \end{aligned} \tag{1.2}$$

Here, $P_{(\theta,\mathbf{X})}$ refers to the joint probability measure of $(\theta, \mathbf{X})$ and $E_{\mathbf{X}}$ refers to taking an average over $\mathbf{X}$. We can rewrite the statement in the Eq. (1.2) from a frequentist angle, noting that

$$\begin{aligned} &P_{(\theta,\mathbf{X})}\{Pos^{-1}(t_1) \le \theta \le Pos^{-1}(t_2)\} \\ &= E_{\theta}[P_{\mathbf{X}|\theta}\{Pos^{-1}\{t_1) \le \theta \le Pos^{-1}(t_2)\}] \\ &= t_2 - t_1. \end{aligned} \tag{1.3}$$

Here the probability $P_{\mathbf{X}|\theta}\{Pos^{-1}(t_1) \le \theta \le Pos^{-1}(t_2)\}$ in the second line of (1.3) is the usual frequentist statement about coverage probability when θ is given. As the parameter θ is random, the E_θ in front is just taking an average over θ. Thus, the coverage statement,

$$P_{(\theta,\mathbf{X})}\{Pos^{-1}(t_1) \le \theta \le Pos^{-1}(t_2)\} = t_2 - t_1, \tag{1.4}$$

is related to both the coverage statement for a credible interval in the Bayes setting and the coverage statement for a confidence interval in the frequentist setting. We refer the coverage statement in (1.2) and (1.3) as a *Bayes Frequentist coverage probability* statement in this paper.

The same Bayes Frequentist coverage probability statement also holds when the nuisance parameter ϕ is included. In particular, as in (1.2), we can show that

$$P_{(\theta,\phi,\mathbf{X})}\{Pos^{-1}(t_1) \le \theta \le Pos^{-1}(t_2)\} = t_2 - t_1, \tag{1.5}$$

from the coverage statement (1.1) of a credible interval in the Bayesian setting. We also can rewrite (1.5) as

$$\begin{aligned} &P_{(\theta,\phi,\mathbf{X})}\{Pos^{-1}(t_1) \le \theta \le Pos^{-1}(t_2)\} \\ &= E_{\theta,\phi}[P_{\mathbf{X}|\theta,\phi}\{Pos^{-1}\{t_1) \le \theta \le Pos^{-1}(t_2)\}] \\ &= t_2 - t_1, \end{aligned} \tag{1.6}$$

which has the same interpretation as Eq. (1.3).

Instead of the Bayes formula, this paper proposes an alternative approach to incorporate prior information with the information from observed data using a recipe developed for combining confidence distributions (CD). In particular, we combine a marginal prior cumulative distribution function on the parameter θ with a confidence distribution on the same parameter derived from the data analysis; for more details, see Section 2. The outcome function from the combination, say $H_c(\theta)$, is a data dependent cumulative distribution function on the θ parameter space. We call such a function a *CD-posterior* function for the parameter θ. We prove in Section 3 that the interval $[H_c^{-1}(t_1), H_c^{-1}(t_2)]$ for θ satisfies the same coverage property as (1.4) and (1.5). Thus, it can substitute the Bayes posterior in forming credible intervals (regions) which are essentially prediction intervals for the parameter θ.

This proposed approach inherits the so call "division of labor" property of frequentist approaches (Efron 1986, Wasserman 2007). Wasserman (2007) states: "The idea that statistical problems do not have to be solved as one coherent whole is anathema to Bayesians but is librating for frequentists". The example that they used to describe the property is estimation of a population quantile θ. While a frequentist approach directly uses the sample quantile to estimate the population quantile, a Bayesian approach requires to assign a prior on the space of all parameters, including both θ and the nuisance parameters ϕ. An estimate of the population quantile θ can then be obtained from the marginal posterior of θ. The inherited property of "division of labor" can help simplify statistical method, especially in situations when it is hard to keep track of nuisance parameters.

The property of "division of labor" can also help ease the computational burden, since the approach only focuses on the marginal prior distribution of the parameter of interest θ and the nuisance parameter part is not involved. Furthermore, the whole development of CD combining involves fairly elementary computations, unless one chooses some tedious method, like a double bootstrap, to form a CD. Compared to full Bayesian analysis with a sampling algorithm involved in it, the computational task involved here is insignificant.

There is an abundant scope for robustness in the construction of a CD-posterior, with respect to prior as well as data. In Section 4, it is established that with a certain choice of combining function the error bound on the CD-posterior cumulative distribution function is $w\varepsilon/(1-\varepsilon)$ when ε is the error bound on the cumulative distribution function of the prior and w is the relative weight chosen for the prior, $0 < w < 1$, in the combining function.

The rest of paper is arranged as follows. Section 2 reviews some recent developments on combining confidence distributions and provides a framework for combining prior information with evidence from observed data. Section 3 provides several examples to demonstrate the proposed approach. Section 4 contains a theoretical development on coverage probabilities. Section 5 establishes an error bound on CD-posterior when there is an error in prior specification. Section 6 contains a simple extension of multiparameter cases. Section 7 contains some further discussions, including a comment on an interesting "amphibious" nature of the proposed methodology — the CD-posterior approach, which is developed under the Bayesian setting assuming parameters are random, also works under the frequentist setting assuming parameters are fixed and nonrandom.

2. The framework

For a given real valued parameter θ, we have a notion of *confidence distribution* (CD) $H_n(\cdot)$, whose t-th quantile is the 100t% upper confidence limit of θ, for all t in (0,1); see, e.g., Efron (1998). This can be restated as follows: $H_n(\cdot)$ is a data-dependent cumulative distribution function such that

$$P(H_n^{-1}(t_1) \leq \theta \leq H_n^{-1}(t_2)) = t_2 - t_1$$

for all $0 < t_1 < t_2 < 1$. This requirement is equivalent to $H_n(\theta) \sim U[0,1]$, uniform distribution on $[0,1]$ when θ is the true parameter value. This is seen by noting that

$$P(\theta \leq H_n^{-1}(t)) = P(H_n(\theta) \leq t).$$

Recently, under the frequentist setting assuming that θ is fixed and non random parameter, Schweder and Hjort (2002), Singh, Xie and Strawderman (2005, 2007) provided a modern version of formal CD definition. Their definition broadens the classical CD notion. In particular, they treat a CD function as a "distribution estimator" of the parameter, in the same sense as one uses a single point (point estimator) or an interval (confidence interval or "interval estimator") to estimate a parameter of interest. Along a similar line, we provide in this article a CD definition for a random parameter θ. Here $\mathcal{X}$ is the sample space for $\mathbf{X}$, Θ is the parameter space for θ and the nuisance parameter is suppressed.

Definition 2.1. *A function $H(\theta) = H(\mathbf{X}, \theta)$ on $\mathcal{X} \times \Theta \to [0,1]$ is called a confidence distribution (CD) for a parameter θ, if $H(\cdot)$ is a continuous cumulative distribution function on Θ for each given sample $\mathbf{X} \in \mathcal{X}$ and $H(\theta)$ has the uniform distribution $U[0,1]$ under the joint distribution of $(\theta, \mathbf{X})$.*

This definition encompasses the case when $H(\mathbf{X}, \theta)$ involves only θ which is the case with the prior distribution function.

A simple example of CD is from the normal mean inference problem with sample $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ from $N(\mu, \sigma^2)$: The basic CD's for μ are $\Phi(\sqrt{n}(\mu - \bar{x})/\sigma)$ when σ is known and $T_{n-1}(\sqrt{n}(\mu - \bar{x})/s)$ when σ is not known, regardless of whether or not (μ, σ^2) has a prior or what the prior is. Here, $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ and $s^2 = (n - 1)^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$ are the sample mean and variance, respectively, and T_{n-1} stands for the cumulative distribution function of student's t-distribution with $n-1$ degrees of freedom. It is clear that a CD defined above satisfies the coverage probability statements (1.2) and (1.4), replacing $Pos^{-1}(\cdot)$ with $H_n^{-1}(\cdot)$.

To define an asymptotic CD, one would simply require $H_n(\theta) \xrightarrow{\mathcal{L}} U[0,1]$. Typically, the bootstrap distribution of a statistics is asymptotic CD. Bootstrapping a studentized statistics yields a second order correct CD. Many non-trivial examples of exact CD can be obtained from regression and ANOVA settings with normal errors. When it exists, we call $h(\theta) = \frac{d}{d\theta} H(\theta)$ a CD density, also known as a confidence density in the literature; see, e.g., Efron (1993).

Assume a marginal prior is available on θ and $H_1(\cdot)$ is the continuous marginal prior cumulative distribution function. Clearly, $H_1(\theta)$ has the uniform distribution $U[0,1]$. According to Definition 2.1, $H_1(\cdot)$ is a CD function for θ. On the other hand, given sample data, assume we can also obtain a CD function, say $H_2(\cdot)$, for the same parameter of interest θ. Now, we have two distribution estimators of θ, one from the prior and one from the sample data. The task is to incorporate the information in these two estimators and construct a combined distribution estimator.

Singh, Xie and Strawderman (2005) proposed a general CD combination recipe using a monotonic mapping function. Xie, Singh and Strawderman (2009) explored weighted combining and developed a general framework for meta-analysis, which encompasses most commonly used meta-analysis approaches. In this this article, we propose to use the same recipe to incorporate $H_1(\cdot)$ the CD function from the prior with $H_2(\cdot)$ the CD function from the sample data. Note that, unless in some special cases, Bayes formula can not always be used with just a marginal prior in presence of other parameters which is the setting of our present discussion; see Section 3 for examples.

Treat $H_1(\cdot)$ and $H_2(\cdot)$ as CD functions derived from two independent sources for a common parameter θ. Let $g_c(u_1, u_2)$ be a function from $[0,1]^2$ to $\mathcal{R}$, which is non-decreasing in each coordinate. Let $G_c(\cdot)$ be the continuous cumulative distribution function of $g_c(U_1, U_2)$; where U_1 and U_2 are independent draws from $U[0,1]$. Then,

$$H_c(x) = G_c(g_c(H_1(x), H_2(x)))$$

is a combined CD for θ, as $H_c(\theta) \sim U[0,1]$. A class of functions worth paying special attention to, is the following:

$$g_c(x_1, x_2) = w_1 F_0^{-1}(x_1) + w_2 F_0^{-1}(x_2) \tag{2.1}$$

where F_0 is a fixed external cumulative distribution function and w_i's are positive fixed constant chosen suitably. Of course, in this case $G_c(\cdot) = F_0(\cdot/w_1) * F_0(\cdot/w_2)$ where $*$ denotes convolution.

There are many choices for the combining function as well as the weights if one chooses to use weights. This can be viewed as an advantage as there is a flexible class of combination methods that are at the disposal of researchers. This can also be a disadvantage as it adds a burden of choice in the procedure and a poor choice of the combination methods can result in loss of efficiency in the combined CD function. The definition of efficiency for a CD function can be found in Schweder and Hjort (2002) and Singh, Xie, Strawderman (2007). Loosely speaking, a CD function that leads to a more precise inference (e.g., narrower confidence interval, etc) is considered as a more efficient CD. Some guidelines are offered in making the choices for combination in Singh, Xie and Stawderman (2005) and Xie, Singh and Strawderman (2009), although further research on this issue is still called for. An encouraging aspect is that the theoretical results developed under the general framework work for all choices, and efficiency loss is not a major issue for many conventional choices of the g_c functions, for instance, the g_c functions corresponding to the commonly used meta-analysis approaches as listed in Xie, Singh and Strawderman (2009).

In our discussion, we can include a nuisance parameter ϕ, which is allowed to be infinite dimensional. Note that, $H_1(\cdot)$ is a marginal prior on θ; whose source may be available expertise, experience and/or past formal/informal trial(s). There may be a joint prior on $\{\theta, \phi\}$, but here we do not require or deal with this unknown joint prior. The most basic example being that of $N(\mu, \sigma^2)$ where $\theta \equiv \mu$ and $\phi \equiv \sigma^2$. In a bivariate normal setting θ could be the coefficient of correlation, while the remaining parameters are ϕ. In a multivariate normal setting θ could be highest of the population means or highest of the populations eigenvalues. For an example of infinite dimensional ϕ, consider the classic case of nonparametric inference problem on $\theta =$ the center of symmetry of an unknown symmetric distribution F; thus ϕ is a symmetric cumulative distribution.

The posterior cumulative distribution of θ Pos$(\cdot)$ is often a CD function. Note that, we typically have

$$P_{\theta|\mathbf{X}}(Pos(\theta) \le t) = t,$$

assuming that $Pos(\cdot)$ is continuous. Hence, unconditionally, we have

$$P_{(\theta,\mathbf{X})}(Pos(\theta) \le t) = t;$$

Thus, the unconditional distribution of $Pos(\theta)$ is $U[0,1]$. By Definition 2.1, Pos$(\cdot)$ is a CD function. Of course, derivation of $Pos(\cdot)$ involves joint prior of $\{\theta,\phi\}$, but ϕ is integrated out from the posterior.

3. Examples

In the most basic case of normal mean problem, the Bayes posterior, when a normal prior is used, coincides with the combined CD H_c, provided a specific combining function g_c is used. This example is described below in detail.

Example 1 (Normal mean problem). Let $Y \sim N(\mu, s^2)$ represent the data. In inferential context Y is a statistic and typically $s^2 = O(1/n)$ with n as the sample size. Let $N(\upsilon,\tau^2)$ be a prior on μ i.e. we assume $\mu \sim N(\upsilon,\tau^2)$. As is well known, the Bayesian posterior distribution of μ is

$$\mu|Y \sim N(aY + (1-a)\upsilon,\ (s^{-2}+\tau^{-2})^{-1}), \tag{3.1}$$

where $a = s^{-2}/(s^{-2}+\tau^{-2})$.

Turning to the CD-posterior, $H_1(\mu) = \Phi((\mu-\upsilon)/\tau), H_2(\mu) = \Phi((\mu-Y)/s)$. Let the g_c in (2.1) be

$$g_c(x_1,x_2) = \frac{1}{\tau}\Phi^{-1}(x_1) + \frac{1}{s}\Phi^{-1}(x_2). \tag{3.2}$$

It leads to $G_c(x) = \Phi(x/(1/\tau^2+1/s^2)^{1/2})$. Therefore,

$$H_c(\mu) = \Phi\left(\frac{(\mu-\upsilon)/\tau^2 + (\mu-Y)/s^2}{(1/\tau^2+1/s^2)^{1/2}}\right).$$

This $H_c(\mu)$ coincides with the cumulative distribution function of the posterior distribution $\mu|Y$ in (3.1).

Let us note the difference between the weights used in the g_c function and the weights in optimal combining of point estimators: the weights used in the optimal combining of point estimators are inversely proportional to variances; here, the weights are inversely proportional to standard deviations. Even if one takes weights $= 1$ in the g_c function, the combined CD will automatically put more weight on the CD with less spread.

The second example relates to a nonparametric location problem. It shows that a CD-posterior can be obtained using just the prior on underlying real valued location parameter, without assuming a parametric density function on the underlying distribution. This is in contrast with a traditional Bayesian approach, which would require the knowledge of the underlying distribution and a prior on the space of the underlying distribution.

Example 2 (Nonparametric location problem). Let μ be the center of symmetry of an unknown symmetric continuous distribution. Suppose there is an independently identically distributed sample available from this population, and also assume that a prior $H_1(\mu)$ is available on μ, but no prior is known on the space of continuous symmetric distributions. A convenient CD in the context would be the following significance function of the class of tests for one-sided hypotheses (see, Fraser, 1991)

$$H_2(x) = \text{p-value of a specified nonparametric test for } H_0: \mu \le x \text{ vs. } H_1: \mu > x.$$

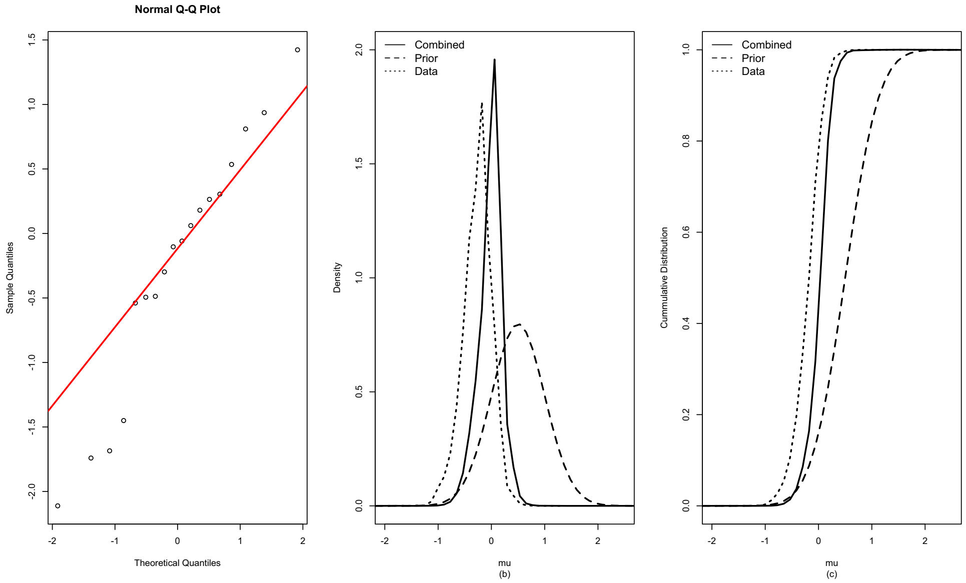

Fig 1. *Illustration of the CD combination approach for incorporating a $N(0.5, 0.5^2)$ prior with a sample from a double exponential distribution in Example 2: (a) qq-plot of the double exponential sample; (b) combination approach illustrated using the densities; (c) combination approach illustrated using the cumulative distributions.*

Suppose, we settle on Sign-Rank Test which has high efficiency for the normal population. It is not a hard exercise to show that such a function $H_2(x)$ is non-decreasing, ranging from 0 to 1. To claim that this $H_2(\cdot)$ is an exact CD, the only obstacle is the fact that the $H_2(\mu)$ is somewhat discrete; the discreteness vanishing as $n \longrightarrow \infty$. Based on this CD function $H_2(\cdot)$ and the prior CD function $H_1(\mu)$, a CD-posterior can be obtained using a recipe described in the previous Section.

A numerical illustration of this example is provided below in Figure 1, where the unknown symmetric distribution is the standard Laplace distribution and the prior for the location parameter is assumed to be $N(0.5, 0.5^2)$. Figure 1 (a) contains a normal quantile plot, clearly indicating this set of $n = 18$ sample values are not from a normal distributed population. The three curves in Figure 1 (b) are, respectively, the prior density curve (the dashed line), the CD density function obtained by taking a numerical derivation of $H_2(\mu)$ (the dotted line) and the density function of the CD-posterior (the solid line). Figure 1 (c) illustrates the same CD combination result as in Figure 1 (b), but presented in the format of cumulative distribution functions. In this numerical example, F_0 used in the combination receipt (2.1) is the cumulative distribution function of the double exponential distribution (the same as the standard Laplace distribution), with no weight $w_1 = w_2 \equiv 1$. Singh et al. (2005) showed that the choice of F_0 offers Bahadur-optimality. Note that, in the example in Figure 1, the the knowledge of the true underlying population (i.e., the standard Laplace distribution) is not used in any way, except that the 18 sample points are generated from that distribution.

The third example is from Xie, Liu, Damaraju and Olson (2009), which is motivated by a consulting project with Johnson and Johnson pharmaceutical company. It is demonstrated that the combining CD method can provide a simple approach to incorporate expert opinions with binomial clinical trial. This example also brings in focus the issue that Bayes formula can not be used with just a marginal prior.

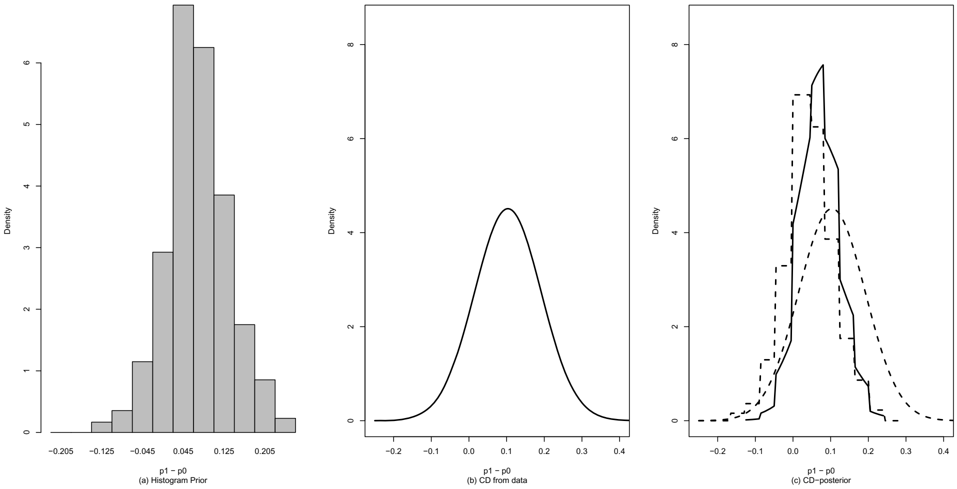

Fig 2. *Illustration of the CD combination approach for Example 3: (a) Histogram obtained from a survey of 11 clinical experts on a new drug treatment (prior information); (b) the asymptotic CD density function of the improvement $\delta = p_1 - p_0$ from the binomial clinical trial; (c) The CD-posterior of the improvement $\delta = p_1 - p_0$ in the density format.*

Example 3 (Binomial clinical trial). Consider a clinical trial that involves two groups: Treatment A and Treatment B, where group B is the control group. Assume that there are n_1 subjects in Treatment A and n_0 in Treatment B. The treatment responses for the two groups are $\{x_{1i}, i = 1, \ldots, n_1\}$ and $\{x_{0i}, i = 1, \ldots, n_0\}$, respectively. Each individual response is a binary outcome and assumed to follow a Bernoulli distribution, i.e., $X_{1i} \sim \text{Bernoulli}\,(p_1)$, for $i = 1, \ldots, n_1$, and $X_{0i} \sim \text{Bernoulli}\,(p_0)$, for $i = 1, \ldots, n_0$. Assume that before the clinical trail a prior of expert opinions on $\delta = p_2 - p_1$ is obtained, which is shown by the histogram in Figure 2 (a); see Xie et al. (2009) for more details. The parameter of interest is $\delta = p_2 - p_1$ and our task is to make inferences on$\delta = p_1 - p_0$, incorporating both the expert opinions and the clinical trial results.

Figure 2 (b) is the density function an asymptotic CD function, $H_2(\delta) = \Phi(\{\delta - \hat{\delta}\}/C_d)$. Here, $\hat{\delta} = \bar{x}_2 - \bar{x}_1 = .1034$ with $\bar{x}_k = n_k^{-1} \sum_{i=1}^{n_k} x_{ki}$, for $k = 0, 1$, and $C_d^2 = \widehat{\text{var}(\hat{\delta})} = .0885^2$; these numerical values are computed from the clinical trial reported in Xie et al. (2009). Let $H_1(\delta)$ be the empirical cumulative distribution function of the histogram of Figure 2(a). Using the g_c in (3.2) with τ^2 being the empirical variance calculated from the histogram and $s = C_d$, we have

$$H_c(\delta) = \Phi\left(\frac{\Phi^{-1}(H_1(\delta))/\tau + \Phi^{-1}(H_2(\delta))/C_d}{(1/\tau^2 + 1/C_d^2)^{\frac{1}{2}}} \right).$$

The density function of the CD-posterior function is plotted in a solid curve in Figure 2(c), together with the dashed curves of $\frac{d}{d\delta} H_1(\delta)$ and $\frac{d}{d\delta} H_2(\delta)$.

In this example, it is not possible to find a "marginal" likelihood of δ, i.e., a conditional density function $f(data|\delta)$. Thus, it is not possible to use Bayes formula to incorporate the marginal expert opinions on δ, without involving an additional parameter. Although one may find an empirical Bayes type solution to such a problem using the concept of estimated likelihood by Boos and Monahan (1986), a full Bayesian solution needs to jointly model δ and an additional parameter or, equivalently, parameters (p_0, p_1); see, e.g., Joseph, du Berger and Belisle (1997) and Xie, et al. (2009). The full Bayesian solution is theoretically sound but computationally demanding. More importantly, Xie et al. (2009) found that, in the case of skewed joint prior distributions, the full Bayesian solution may sometimes generate a counterintuitive "outlying posterior phenomenon" in which the marginal posterior of δ is more extreme than both its prior and the likelihood evidence. Further details and discussions on this paradoxical phenomenon can be found in Xie et al. (2009).

4. The coverage theorem

Suppose that our sample data are obtained by the following hierarchical process underlying the Bayesian philosophy,

- First, parameters (θ, ϕ) are drawn from a prior distribution $\pi(\theta, \phi)$
- Then, conditional on the realization of the parameter set (θ, ϕ), sample data $\mathbf{X}$ are drawn from a conditional distribution $F(\mathbf{X}|\theta, \phi)$.

We assume that marginal prior cumulative distribution function $H_1(t) = \int_{-\infty}^{t} \int \pi(\theta, \phi) d\phi d\theta$ of θ is continuous. The CD function constructed from the sample $H_2(\theta) = H_2(\mathbf{X}, \theta)$ satisfies

$$H_2(\theta)|\theta \sim U(0,1) \tag{4.1}$$

as a function of random sample $\mathbf{X}$, which is the requirement for a CD in a Frequentist setting; See, e.g., Singh, Xie and Strawderman (2007) for further details. From (4.1), $H_2(\theta)$ is also $U[0,1]$ under the joint distribution of $(\theta, \mathbf{X})$ or $(\theta, \phi, \mathbf{X})$. Thus, $H_2(\theta)$ is also a CD under the Definition 2.1 for a random θ. We combine $H_1(\cdot)$ from the prior and $H_2(\cdot)$ from the sample data using the recipe of combining confidence distributions described in the previous section. The following theorem asserts that the CD-posterior function $H_c(\cdot)$ offers the coverage probability stated in (1.4) and (1.5). It entails that the interval $[H_c^{-1}(.025), H_c^{-1}(.975)]$ has 95% of probability to cover θ, a statement which takes into account the randomness of the vector $\{\theta, \phi\}$.

Theorem 4.1. *Assume that $H_2(\theta)$ obtained from the sample satisfies (4.1). Under the joint distribution of either $\{\theta, \mathbf{X}\}$ or $\{\theta, \phi, \mathbf{X}\}$, $H_c(\theta) \sim U[0,1]$. Hence, $H_c(\theta)$ is a CD function and we have the following statement on coverage:*

$$P_{(\theta,\mathbf{X})}\{H_c^{-1}(t_1) \le \theta \le H_c^{-1}(t_2)\} = P_{(\theta,\phi,\mathbf{X})}\{H_c^{-1}(t_1) \le \theta \le H_c^{-1}(t_2)\} = t_2 - t_1$$

for all $0 < t_1 < t_2 < 1$.

Proof of Theorem. Let us start out by noting that $H_1(\theta)$ and $H_2(\theta)$ both have $U[0,1]$ distribution. If we show that $H_1(\theta)$ and $H_2(\theta)$ are independent, the claim will follow, noting the fact that

$$H_c(\theta) = G_c(g_c(H_1(\theta), H_2(\theta)))$$

where G_c is the continuous cumulative distribution function of $g_c(U_1, U_2)$ with U_1, U_2 as two independent draws from $U[0,1]$.

The key task is to show that $H_1(\theta)$ and $H_2(\theta)$ are independent, even though they both share a common random variable θ. Towards this end, we argue as follows: Fixing $H_1(\theta) \Leftrightarrow$ fixing θ (with probability 1 if H_1 is not strictly increasing). Let

$H_1^{-1}(\cdot)$ denote the left continuous inverse. Now, for any fixed t and s, $0 < t < 1$, $0 < s < 1$, and $\theta_s = H_1^{-1}(s)$, we have

$$P_{(\theta,\mathbf{X})|H_1(\theta)=s}\{H_2(\theta) \leq t\} = P_{\mathbf{X}|\theta=\theta_s}\{H_2(\theta_s) \leq t\} = t,$$

in absence of a nuisance parameter. Similarly, in presence of a nuisance parameter ϕ, we have

$$\begin{aligned} P_{(\theta,\phi,\mathbf{X})|H_1(\theta)=s}\{H_2(\theta) \leq t\} &= P_{(\phi,\mathbf{X})|\theta=\theta_s}\{H_2(\theta_s) \leq t\} \\ &= E_{\phi|\theta=\theta_s} P_{\mathbf{X}|\theta=\theta_s,\phi}\{H_2(\theta_s) \leq t\} \\ &= E_{\phi|\theta=\theta_s}(t) \\ &= t. \end{aligned}$$

Thus, the proof is complete. □

5. On prior robustness

A prior distribution by its very nature is subject to some misspecification. It is desirable to have a robust procedure that is insensitive to error in a prior up to a certain degree. Let us suppose, there is an ϵ error in prior cumulative distribution function H_1, i.e.

$$\sup |H_1(x) - H_1^*(x)| = \varepsilon$$

where H_1^* is the true prior. It is well known that the standard Bayesian posterior can be perturbed to an arbitrary amount under such an ε error. For reader's convenience, we include an example here. In Example 1 of Section 3, suppose that H_1^* is simply $N(\upsilon, \tau^2)$, truncated at τz_t in the upper side, where z_t is the $(1-t)$th quantile of the standard normal distribution. Thus,

$$\begin{aligned} H_1^*(x) &= (1-t)^{-1} H_1(x) \text{ for } x \leq \tau z_t \\ &= 1, \text{ otherwise.} \end{aligned}$$

Under this setting with a truncated normal prior H_1^*, the Bayes posterior, say $Pos^*(x)$, will place no mass above τz_t. On the other hand, with respect to a non-truncated normal prior H_1, the posterior, say $Pos(x)$, is normal with mean $aY + (1-a)\upsilon$ which can take any value on the real line. One may adjust t to make the sup difference of the priors H_1 and H_1^* arbitrarily small; however, the sup difference between corresponding posteriors $Pos(x)$ and $Pos^*(x)$ can be arbitrarily close to 1.

Let us turn our attention to the CD-posterior. We offer here a specific combining function g_c which yields a very promising bound on the output CD function H_c. Consider the combining function

$$g_c(x_1, x_2) = wx_1 + (1-w)x_2, \quad 0 < w < 1. \tag{5.1}$$

We have the following theorem.

Theorem 5.1. *Suppose there is an error bounded by ϵ in the prior H_1 and an error bounded by δ in the data-based CD, H_2. If we use the g_c function in (5.1) in our combination, the resulting error in H_c is bounded by*

$$\frac{w}{1-w}\epsilon + \delta \quad \textit{if} \quad 0 < w \leq 1/2$$

$$\epsilon + \frac{1-w}{w}\delta \quad \textit{if} \quad \frac{1}{2} \leq w < 1$$

Clearly, $w < 1/2$ is more realistic, since the prior H_1 would typically be a lot more spread out than the CD H_2. We are unable to provide any concrete choice of w, but a reasonable approach would be to choose w by minimizing the spread of $H_c(\cdot)$ (by some measure of scale). For small w, the bound is just a fraction of ϵ. For a starter, the choice of weights inversely proportional to corresponding standard deviations, seems sensible as well. See, also, Xie, Singh and Strawderman (2009), in which a g_c function similar to (5.1) is used to develop a robust meta-analysis procedure under the frequentist setting.

Proof of the Theorem. The proof requires the density function, say $h(x)$, of $wU_1 + (1-w)U_2$, where U_1 and U_2 are independent $U[0,1]$ random variables. This density $h(x)$ can be yielded by a standard derivation. First, consider the case when $0 < w \leq \frac{1}{2}$. In this case, for $0 \leq x \leq 1-w$, we have

$$\begin{aligned} h(x) &= \frac{x}{w(1-w)} \ , \quad \text{for } \ 0 \leq x \leq w \\ &= \frac{1}{1-w} \ , \qquad \text{for } \ w \leq x \leq 1-w. \end{aligned}$$

For $1 \geq x \geq 1-w$, we have

$$h(x) = h(1-x) \ ,$$

In the case when $w > \frac{1}{2}$, the density $h(x)$ can be obtained by simply replacing w with $1-w$ throughout the expression of the above $h(x)$.

In the case when $w \leq \frac{1}{2}$ the density is bounded by $(1-w)^{-1}$. The ϵ and δ perturbations in H_1 and H_2 respectively will perturb $g_c(H_1(x), H_2(x))$ by $w\epsilon + (1-w)\delta$; which will perturb the function $H_c(x) = G_c(g_c(H_1,(x), H_2(x)))$ by $\frac{w}{1-w}\epsilon + \delta$, at most. In the case when $w > \frac{1}{2}$, the argument is similar. This proves the theorem. □

We conclude the section with the remark that if the above g_c is replaced by normal-based combining function i.e.

$$w\Phi^{-1}(x_1) + (1-w)\Phi^{-1}(x_2)$$

the resulting CD-posterior H_c will exhibit behavior just like the Bayes Posterior, at least for normal samples.

6. Multiparameter extension

The foregoing technique of combining does not seem to have any natural extension to $\mathcal{R}^k$ i.e. to the case when one is attempting to combine joint prior on k parameters with an inference function like a CD. However, there is a substitute available which can serve the same purpose. The substitute to CD is a p-function of point-hypotheses, which is combined with a suitable function derived from a prior cumulative distribution function, using the same combining recipe. The details are as follows.

For a parameter $\boldsymbol{\theta} \in \mathcal{R}^k$, define

$$p(\mathbf{x}) = \text{p-value of some specified test for testing } H_0 : \theta = \mathbf{x} \text{ vs } H_1 : \theta \neq \mathbf{x}.$$

Note the difference between this H_0 and the H_0 used in the nonparametric example (Example 2) given earlier. This significance function (p-value function) can be used to define a confidence region at 100t% level as

$$C_t = \{\mathbf{x} \ \epsilon\mathcal{R}^k \ \text{ s.t. } \ p(\mathbf{x}) \geq 1-t\}.$$

Since $p(\boldsymbol{\theta}) \sim U[0,1]$,

$$P_{\mathbf{X}}(\boldsymbol{\theta} \in C_t) = P(p(\boldsymbol{\theta}) \geq 1-t) = t.$$

Clearly, the same statement holds with respect to $P_{\mathbf{X},\boldsymbol{\theta}}$ or $P_{\mathbf{X},\boldsymbol{\theta},\phi}$.

A point estimator for $\boldsymbol{\theta}$, using the p-value function could be defined as

$$\hat{\boldsymbol{\theta}} = \arg\max p(\mathbf{x}),$$

A test for $H_0 : \theta \in S$ vs $H_1 : \theta \notin S$, when S is a region, could be reasonably carried out at a level α by rejecting H_0 when

$$C_{1-\alpha} \cap S = \emptyset$$

Note that, such a test has type one error $\leq \alpha$ because, under $H_0, \theta \in S$, the test rejects $H_0 \Rightarrow p(\boldsymbol{\theta}) < \alpha$ which has probability α. Thus, all three types of frequentist inference can be carried out with the p-value function of point hypotheses.

Consider now a prior distribution for $\boldsymbol{\theta}$, which is a cumulative distribution function G on $\mathcal{R}^k$. One needs to define a function $c(\cdot)$ based on G which is compatible with $p(\cdot)$. Such a function should take values in [0,1] and have the following additional properties:

1. $c(\mathbf{x})$ should measure centrality of $\mathbf{x}$ with respect to the distribution G.
2. $c(\boldsymbol{\theta}) \sim U[0,1]$ when $\boldsymbol{\theta} \sim$ G.

The concept of data-depth comes to the rescue! Define

$$c(\mathbf{x}) = P_G\{\mathbf{y} \in \mathcal{R}^k \text{ such that } D_G(\mathbf{x}) \geq D_G(\mathbf{y})\}$$

Thus $c(\mathbf{x}) = 1$ when $\mathbf{x}$ maximizes $D_G(\mathbf{x})$ over $\mathcal{R}^k$; also $c(\mathbf{x})$ approaches towards 0, as $D_G(\mathbf{x})$ moves towards its minimum values (i.e. $\mathbf{x}$ towards the outskirts of G). The function $c(\mathbf{x})$ simply measures the centrality of $\mathbf{x}$ by evaluating the probability content of the portion less deep than $\mathbf{x}$ itself. Some of the most popular notions of data-depth being Tukey's depth, Mahalanobis depth (reciprocal of {1 + Mahalanobis distance} from the center), Liu's simplicial depth; see, e.g., Liu, Parelius and Singh (1999). It is well documented in the literature on data-depth that the centrality random variable $c(\boldsymbol{\theta}) \sim U[0,1]$ when $\theta \sim G$, provided the distribution of $D_G(\boldsymbol{\theta})$ is continuous, (see Liu and Singh, 1993). The function $c(\mathbf{x})$ derived from the prior distribution possesses the characteristics of a significance function (p-value function).

The combining recipe remains unaltered:

$$p_c(\mathbf{x}) = G_c(g_c(c(\mathbf{x}), p(\mathbf{x})))$$

where $g_c(\cdot)$ is the combining function, G_c is the cumulative distribution function of $g_c(U_1, U_2)$ as in Section 2. Evidently, $p_c(\boldsymbol{\theta}) \sim U[0,1]$ and

$$C_c(\alpha) = \{\mathbf{x} : p_c(\mathbf{x}) \geq 1-\alpha\}$$

can serve as combined confidence region at 100t% level. Combined point estimator would be defined a

$$\hat{\theta}_c = \arg\max p_c(\mathbf{x})$$

The issue of choosing the combining function $g_c(\cdot)$ remains unexplored.

We close this section by providing an example of the centrality function $c(\cdot)$ in the most basic case when G is multivariate normal.

Example 4. Suppose $\boldsymbol{\mu} \sim N(\boldsymbol{\upsilon}, \Sigma)$, where $\boldsymbol{\mu}$ is a vector valued $(k \times 1)$ parameter. The $N(\cdot, \cdot)$ notation has its standard meaning: $\boldsymbol{\upsilon}$ is the mean vector and Σ is the dispersion matrix. Mahalanobis distance of a point $\mathbf{x}$ from $\boldsymbol{\upsilon}$ is

$$(\mathbf{x} - \boldsymbol{\upsilon})^1 \Sigma^{-1} (\mathbf{x} - \boldsymbol{\upsilon})$$

Let $c(\cdot)$ be the probability content of all points $\mathbf{x}$ which are at a higher Mahalanobis distance than $\mathbf{x}$ (i.e. have lower data-depth). Then it follows that

$$c(\mathbf{x}) = P(\chi_k^2 \geq (\mathbf{x} - \boldsymbol{\upsilon})^1 \Sigma^{-1} (\mathbf{x} - \boldsymbol{\upsilon}))$$

where χ_k^2 is a random variable having the standard chi-squire distribution with k d.f. The claim is based on the fact that $(\boldsymbol{\mu} - \boldsymbol{\upsilon})^1 \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\upsilon})$ has chi-squire (k) distribution. As a matter of fact the formula for the centrality function $c(\cdot)$ remains the same, as long as one is using an affine invariant depth (including Tukey's depth, simplicial depth).

7. Discussions and conclusions

The article attempts to provide an alternative procedure to synthesize prior information on a parameter with the inference emerging from data under a Bayesian paradigm. By simply combining prior distribution with data based CD, this procedure produces a function, called a CD-posterior, which is typically the same as the usual Bayesian posterior in the basic normal case but different, otherwise. Interestingly, a confidence/credible interval derived from a CD-posterior offers the same statement of coverage probability of the credible interval as a Bayesian approach. A key advantage that this methodology has is that the prior distribution is required only on the parameter of interest, not on the whole parameter-vector appearing in the likelihood function. Also, the proposed approach is computationally simple and the saving in computational effort could just be phenomenal, especially compared to some full blown Bayesian analysis which needs to use MCMC algorithms. The proposed methodology also includes a class of robust combining procedures in which an error bound is established on the CD-posterior when there is an error in prior specification.

The key advantage that this methodology can directly zoom in and focus on the parameter of interest can have further implications. In order to carry out Bayesian analysis, a statistician is ideally required to come up with a joint prior distribution of the parameter being studied and the nuisance parameters involved in the likelihood function. Often times such "full priors" are simply choice of statistician's convenience which makes the analysis possible. In the proposed approach, one would simply take the (marginal) prior on the parameter being studied and combine it with a CD derived using one of many tools available in a frequentist's tool box such as asympotics, bootstrap, parametric, nonparametric, semi-parametric methods. The burden of joint prior is taken out, though one may end up using an asymptotic CD, in many cases. Among the asymptotic CD's one often has a choice of having it correct up to a desired asymptotic order. Since the prior is needed only on the parameter being studied, a greater degree of truthfulness and hence accuracy is expected in the construction of the prior.

A specific point which should be brought to light here is the fact that this methodology is flexible in terms of treating the parameter θ as "fixed" as in the frequentist domain or "random" as in the Bayesian domain. A frequentist who refuses to accept the randomness of the parameter, may accept the knowledge based prior as a rough and dirty CD on the parameter by vaguely defining a sample space of "past experiments or experiences". With such an understanding, this CD-posterior is simply a combined CD and it offers exact frequentists' inference which incorporates past knowledge with the data; see, Xie et al. (2009). Thus, combining prior with data based CD has got an amphibious character. Another related article is Bickel (2006), which used a CD combination approach to incorporate expert opinions to a normal clinical trial. Bickel (2006) used an objective Bayesian argument to justify his treatment of the prior information from expert opinions as a CD function. His numerical results also illustrate the tremendous saving in computational effort and it clearly demonstrates the potential of a CD combination approach in Pharmaceutical practice.

It would be only fair to include some comparative advantages of the Bayesian analysis as well. Once a statistician settles on a prior and a likelihood using whatever principle or mechanism, the Bayesian analysis is unique. However, that is not the case in CD based analysis. There is huge amount of choices for the combining function as well as the weights if one chooses to use weights. In the text, some guidelines are offered in making these choices but further research on this issue is called for. Also, once a Bayes posterior is derived, there is a Bayesian inference available in principle on any quantity of interest attached to the model, but the proposed CD-posterior targets only a chosen parameter. Besides regular full blown Bayesian analysis, there are clever short-cuts that have been proposed in the literature; see Boos and Monahan (1986), among others. These approximate Bayes methods do offer an appreciable degree of simplification in many cases, though typically such methods make simplifying assumptions and are valid only asymptotically; unlike the CD posterior which offers an exact coverage theory.

References

Bickel D.R. (2006) Incorporation expert knowledge into frequentist inference by combining generalized confidence distributions. Unpublished manuscript.

Boos, D.D. and Monahan, J.F. (1986). Bootstrap methods using prior information. *Biometrika*, 73, 77–83.

Efron, B. (1986). Why isn't everybody a Bayesian? *The American Statistician*, 40(1), 1–5.

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika*, 80, 3–26.

Efron, B. (1998). Fisher in 21st Century (with discussion) *Stat. Scie.*, 13, 95–122.

Fraser, D.A.S. (1991). Statistical inference: Likelihood to significance. *Journal of the American Statistical Association*, 86, 258–265.

Joseph, L., du Berger R., and Belisle P. (1997). Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine*, 16, 769–781.

Liu, R.Y. and Singh, K. (1993). A quality index based on data-depth and a multivariate rank test. *Journal of the American Statistical Association*, 88, 257–260.

Liu, R.Y., Parelius, J. and Singh, K. (1999). Multivariate analysis by data-depth: Descriptive statistics, graphics and inference. *Ann. Stat.*, 27, 783–856 (with discussions).

Schweder, T. and Hjort, N.L. (2002). Confidence and Likelihood. *Scan. J. Statist.*, 29, 309–332.

Singh, K., Xie, M. and Strawderman, W.E. (2005). Combining information from independent sources through confidence distribution. *Ann. Stat.*, 33, 159–183.

Singh, K., Xie, M. and Strawderman, W.E. (2007). Confidence distributions - Distribution estimator of a parameter. in *Complex Datasets and Inverse Problems.* IMS Lecture Notes-Monograph Series, No. 54, (R. Liu, et al., Eds.), 132–150. Festschrift in memory of Yehuda Vardi. IMS Lecture Notes Series.

Wasserman, L. (2007). Why isn't everyone a Bayesian? in *The Science of Bradley Efron.* (C.R., Morris and R. Tibshirani, Eds.), 260–261. Springer.

Xie, M., Singh, K. and Strawderman, W.E. (2009). Confidence distributions and a unifying framework for meta-analysis. Technical Report. Department of Statistics, Rutgers University. Submitted for publication.

Xie, M., Liu, R.Y., Damaraju, C.V. and Olson, W.H. (2009). Incorporating expert opinions in the analysis of binomial clinical trials. Technical Report. Department of Statistics, Rutgers University. Submitted for publication.