

A predictive look at Bayesian Bandits

Un approccio predittivo ai Bandit Bayesiani

Pierpaolo Brutti, Fulvio De Santis and Stefania Gubbiotti

Abstract A multi-armed bandit problem models an agent that simultaneously attempts to acquire new information (*exploration*) and optimizes the decisions based on existing knowledge (*exploitation*). In clinical trials, this framework applies to Bayesian multi-armed randomized adaptive designs. The allocation rule of experimental units involves the posterior probability of each treatment being the best. The trade-off between gain in information and selection of the most promising treatment is modulated by a quantity γ , typically prefixed or linearly increasing with accumulating sample size. We propose a predictive criterion for selecting γ that also allows its progressive reassessment based on interim analyses data.

Abstract *Un multi-armed bandit è un esperimento sequenziale in cui un soggetto, per scegliere tra più azioni alternative, tenta simultaneamente di acquisire nuova informazione (“exploration”) e di ottimizzare le decisioni sulla base dei dati disponibili (“exploitation”). Nelle prove cliniche, lo schema si applica ai disegni Bayesiani randomizzati adattivi in cui si confrontano due o più trattamenti. Le unità sperimentali vengono assegnate ai diversi gruppi mediante regole di allocazione basate sulle probabilità a posteriori che ciascun trattamento sia il migliore. Il compromesso tra le contrapposte esigenze di esplorare e selezionare con maggior frequenza il trattamento più promettente è modulato da una quantità γ che tipicamente è prefissata o cresce linearmente con il numero di osservazioni. La nostra proposta prevede invece l’adozione di opportuni criteri predittivi per la scelta iniziale di γ e il suo progressivo aggiornamento sulla base dei dati effettivamente osservati ad interim.*

Key words: Bayesian response–adaptive designs, Exploration/exploitation trade–off, Predictive two–priors approach, Randomized probability matching.

Pierpaolo Brutti

Dip. Scienze Statistiche, Sapienza Università di Roma e-mail: pierpaolo.brutti@uniroma1.it

Fulvio De Santis

Dip. Scienze Statistiche, Sapienza Università di Roma e-mail: fulvio.desantis@uniroma1.it

Stefania Gubbiotti

Dip. Scienze Statistiche, Sapienza Università di Roma e-mail: stefania.gubbiotti@uniroma1.it

1 Introduction

A *multi-armed bandit* problem is a sequential allocation experiment aimed at obtaining the largest possible *reward* by optimally allocating unit resources to alternative treatment arms (for an overview on this topic see among others [5]). The metaphoric term *bandit* alludes to a *one-armed bandit* which colloquially indicates a slot machine. In our setup the payoff associated to an action is simply the response under a given treatment arm and the overall objective is to maximize the number of successes obtained in a sequence of allocations. At each stage the decision-maker must simultaneously respond to a twofold purpose: (i) *exploiting* the treatment that appears to be the most promising in light of the current available knowledge, and (ii) keep *exploring* all arms in search of better treatments that momentarily appear to be inferior, due to limited experimental information (see [1] for discussion). An overview of several techniques used to manage the multi-armed bandit problem is provided by [3], with special emphasis on the so-called *randomized probability matching* and on its implementation in a Bayesian framework which naturally allows to handle the trade-off between exploration and exploitation.

Although multi-armed bandits are currently used in many fields, ranging from website optimization to packet routing, clinical trials still represent the most classical application context (see [2]). In this area the typical Bayesian adaptive randomized (BAR) design for binary outcomes is actually a *binomial bandit* (see [3]).

This paper is organized as follows. In the next section we briefly sketch the basic idea underlying BAR, first proposed by Thompson [8] and subsequently developed in several papers and applied in actual trials. Then in Section 3.2 we suggest predictive criteria for selecting a suitable tuning parameter for the trade-off between exploration and exploitation. This predictive approach is based on the use of two distinct prior distributions in the sense explained in Section 3.1. Finally Section 5 contains some concluding remarks and open problems.

2 Bayesian adaptive randomization

An adaptive clinical trial includes a prospectively planned opportunity for modification of one or more specified aspects of the study design. In this paper we specifically focus on adaptive modifications of treatment randomization probabilities, according to the accumulating observed responses at interim analyses. The Bayesian approach has provided the ideal framework to develop this idea, thanks to the straightforward iterative application of Bayes theorem for knowledge updating. For a comprehensive review on Bayesian adaptive designs see [4].

Here we consider the simplest case of a comparative clinical trial with binary outcomes under two competing treatments. In this scheme each patient is enrolled sequentially and is randomly assigned to one treatment arm based on a suitable allocation rule that will be discussed later on. Once a generic patient i is assigned to arm k with $k = 1, 2$ indicating the new and the standard treatment respectively, let

$X_{k,i}$ denote her binary response to treatment k , which follows a Bernoulli distribution with unknown response probability θ_k . Moreover, let $Y_{k,n_k} = \sum_{i=1}^{n_k} X_{k,i}$ denote the random number of successes out of n_k patients under treatment k . For the current sample size $n = n_1 + n_2$, the observed data will be indicated by $\mathbf{y}_n = (y_{1,n_1}, y_{2,n_2})$.

We assume that available prior information on the unknown parameters θ_1 and θ_2 is formalized by a prior distribution of independence $\pi(\theta_1, \theta_2) = \pi(\theta_1)\pi(\theta_2)$, where each $\pi(\theta_k)$ is a conjugate Beta prior density of parameters (α_k, β_k) to be updated given the data y_{k,n_k} . However, it is worth noticing that both these assumptions, although realistic and widely used in many application contexts, are not strictly required in the following procedure.

Back to the randomization mechanism, the basic ingredient to define the allocation rule is the posterior probability of superiority of the new treatment with respect to the standard one, namely

$$\mathbb{P}_\pi(\theta_1 > \theta_2 | \mathbf{y}_n). \quad (1)$$

As argued in [7], in order to avoid the undesirable variability obtained when using (1) *tout court* as probability of allocation to arm 1, it is common practice to consider the following stabilizing transformation:

$$\frac{\mathbb{P}_\pi(\theta_1 > \theta_2 | \mathbf{y}_n)^\gamma}{\mathbb{P}_\pi(\theta_1 > \theta_2 | \mathbf{y}_n)^\gamma + \mathbb{P}_\pi(\theta_1 < \theta_2 | \mathbf{y}_n)^\gamma}, \quad (2)$$

where γ is a positive quantity that modulates the trade-off between exploration and exploitation. It is straightforward to see that $\gamma = 0$ leads to equal allocation, while for $\gamma = 1$ we retrieve probability (1). In practice, in [7] a value of $\gamma = 1/2$ is recommended, based on empirical experience. Alternatively, it is possible to set γ as a linear function of n/N , where n is the current sample size and N is the prefixed maximum dimension of the trial. In [9] a practical method for the selection of γ is outlined. The Authors focus on a function of the form $\gamma(n) = a(n/N)^b$ depending on two parameters a and b and they suggest to find optimal values of a and b in terms of the operating characteristics of the adaptive design under realistic scenarios. In this work, we propose instead to adopt a predictive approach for selecting γ , that will be detailed in the next section.

3 Our proposal

Before describing our proposal, a brief premise is necessary to introduce the distinction between two prior distributions involved in the criteria that will be described in Section 3.2.

3.1 Predictive two-priors approach

In our predictive setup, following several previous contributions (see, among others, [10]), we consider two distinct prior distributions. Specifically,

- π_A (*analysis prior*) is employed to determine the posterior distribution and, therefore, the posterior probability (1), involved in the allocation mechanism of experimental units as well as in the inferential analysis;
- π_D (*design prior*) describes a conjectured scenario (for example superiority of the new treatment) and induces the marginal distribution of the data m_D upon which the proposed predictive criteria rely on.

Notice that the marginal distribution m_D exists only if π_D is a proper distribution, whereas π_A can also be a noninformative improper prior.

3.2 Predictive criteria to select γ

In this section we introduce two alternative predictive criteria to optimize the choice of $\gamma = \gamma(n)$ that will be then used to modulate the allocation probabilities for randomizing the next n experimental units. Reasonably, under a given scenario depicted by π_D , our goal is to select γ in such a way that (i) the best arm is privileged during the allocation process and (ii) it is likely to identify with satisfying accuracy the most effective treatment in terms of its posterior probability of superiority. To this end, the optimal value γ^* is selected to maximize a suitable summary of the predictive distribution of the posterior probability in (1), such as, for example, the predictive expectation

$$\mathbb{E}_{\pi_D} [\mathbb{P}_{\pi_A} (\theta_1 > \theta_2 | \mathbf{y}_n)] \quad (3)$$

or the predictive probability

$$\mathbb{P}_{\pi_D} [\mathbb{P}_{\pi_A} (\theta_1 > \theta_2 | \mathbf{y}_n) > \lambda^U] - \mathbb{P}_{\pi_D} [\mathbb{P}_{\pi_A} (\theta_1 > \theta_2 | \mathbf{y}_n) < \lambda^L] \quad (4)$$

where λ^U and λ^L are prefixed cutoffs (e.g. $\lambda^U = 0.9$, $\lambda^L = 0.1$), introduced to control the tail behavior of this distribution.

In addition, in the ongoing trial a certain number of interim analyses is planned. At each intermediate stage j the prior distributions π_A is updated given the actually observed data, and the chosen criterion is repeatedly applied, with $n = n^j$ denoting the sample size of the next interim analysis. In particular, notice that it is also possible to update π_D , in the spirit of the *Bayesian interim power* described in [6]: this allows to reassess the optimal value γ^* in light of the actual progress of the trial, even if the real interim data turn out to be in contrast with the initial prefixed scenario.

4 Simulation study

In order to assess the performance of the predictive criteria introduced in the previous section, we have conducted an extensive simulation study, referring to a previous example by [7]. Figure 1 shows a relevant example under the following setup:

- the true underlying response rates are 0.45 and 0.25 for arm 1 and 2 respectively;
- the analysis prior is neutral, that is $\pi_A(\theta_1) = \pi_A(\theta_2) = \text{Beta}(\cdot|1.25, 3.75)$;
- the design prior is slightly optimistic towards the new treatment, with $\pi_D(\theta_1) = \text{Beta}(\theta_1|3, 7)$ and $\pi_D(\theta_2) = \text{Beta}(\theta_2|2.5, 7.5)$.

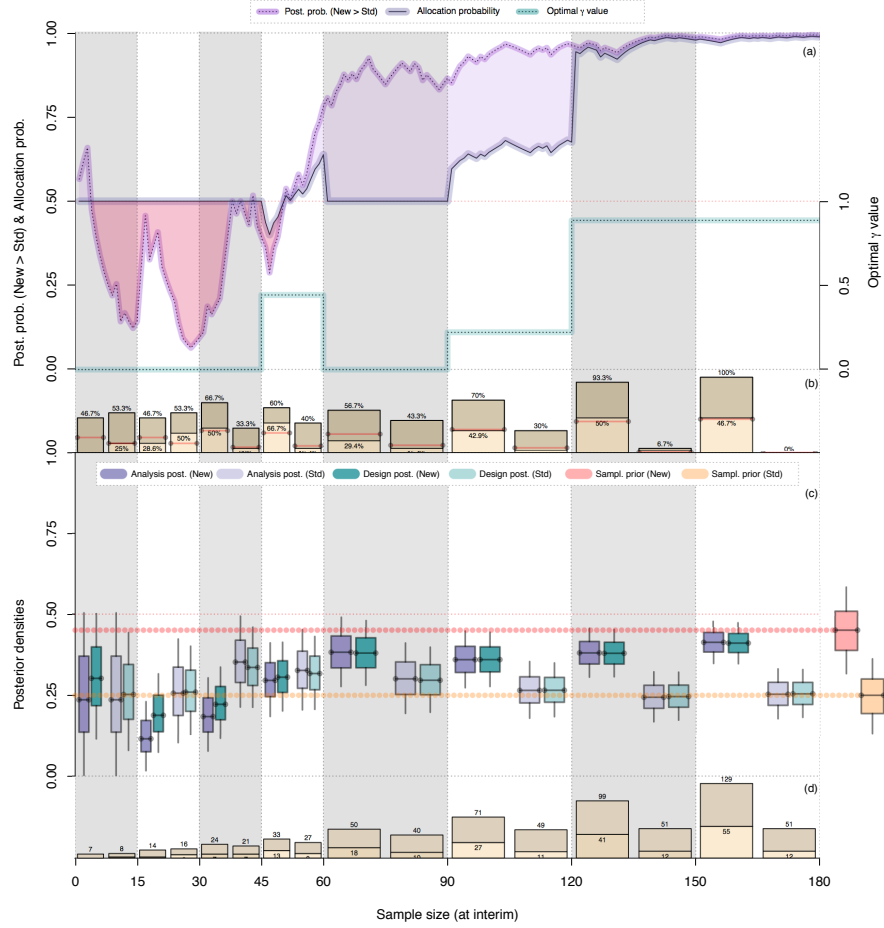


Fig. 1 Panels description (from top to bottom): **(a)** Progressive updating of the (analysis) posterior probability that $\theta_1 > \theta_2$ and the corresponding probability of allocation to arm 1 with respect to the sample size; optimal choices of γ^* at each interim analysis; **(b)** Actual allocation proportions to arms 1 and 2 with the corresponding observed response proportions at each interim analysis; **(c)** Progressive updating of analysis and design priors given the interim data; **(d)** Cumulated sample sizes and absolute number of successes under 1 and 2 at each interim analysis.

At the beginning of the trial, during the first two interim analyses, the control arm seems to perform better than the innovative treatment. This is in contrast with the expectations of the experimenter summarized by the design scenario. Hence, even if the posterior probability (1) tends to be smaller than 0.5, apparently privileging the allocation in arm 2, the selected value of $\gamma^* = 0$ actually leads to equal allocation, allowing for a further exploration of both arms. In the subsequent interim analyses, since the actual data progressively appear to become more favorable towards treatment 1, γ^* starts to increase and finally allows the allocation probability to be very close to 1, which implies that almost all new patients are assigned to arm 1.

5 Concluding remarks

In this paper we consider predictive criteria for selecting the exploration/exploitation parameter γ involved in Bayesian adaptive randomized designs. The proposed approach is extremely flexible and in simulation studies it appears to be promising. In future research we would like to further compare our procedure with existing methods, possibly by introducing suitable stopping rules that allow early termination for either efficacy or futility. Finally, we will investigate the problem of covariate-adjustment, already addressed for instance by [11].

References

1. Cohen J.D., McClure S.M., Yu A.J.: Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Phil. Trans. R. Soc. B* **362**, 933–942 (2007).
2. Hu F, Rosenberger W. *The Theory of Response-Adaptive Randomization in Clinical Trials*, Vol. 525. Wiley-Interscience: Hoboken NJ. (2006).
3. Scott S.L.: A modern Bayesian look at the multi-armed bandit. *Appl. Stochastic Models Bus. Ind.* **26**, 639–658 (2010).
4. Berry S.M., Carlin B.P., Lee J.J., Muller P.: *Bayesian Adaptive Methods for Clinical Trials*. CRC Press (2010).
5. Bubeck S., Cesa-Bianchi N.: Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning* **5**(1), 1–122 (2012).
6. Spiegelhalter, D.J., Abrams, K.R. and Myles, J.P.: *Bayesian approaches to clinical trials and health-care evaluation*. Wiley (2004).
7. Thall P.F., Wathen J.K., Practical Bayesian adaptive randomisation in clinical trials. *Eur. J. Cancer*. **43**, 859–866 (2007).
8. Thompson W.R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*. **25**, 285–294 (1933).
9. Trippa L., Lee E.Q., Wen P.Y., Batchelor T.T., Cloughesy T., Parmigiani G., Alexander B.M.: Bayesian Adaptive Randomized Trial Design for Patients With Recurrent Glioblastom. *J. Clin. Oncol.* **30**(26), 3258–3263 (2012).
10. Wang, F., and Gelfand, A.E.: A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Stat. Sci.* **17**(2), 193–208 (2002).
11. Yuan Y., Huang X., Liu S.: A Bayesian response-adaptive covariate-balanced randomization design with application to a leukemia clinical trial. *Statist. Med.* **30**, 1218–1229 (2011).