

# Bayesian two-sample tests

Karsten M. Borgwardt<sup>1</sup> and Zoubin Ghahramani<sup>2</sup>  
<sup>1</sup>Max-Planck-Institutes Tübingen, <sup>2</sup>University of Cambridge

June 22, 2009

## Abstract

In this paper, we present two classes of Bayesian approaches to the two-sample problem. Our first class of methods extends the Bayesian t-test to include all parametric models in the exponential family and their conjugate priors. Our second class of methods uses Dirichlet process mixtures (DPM) of such conjugate-exponential distributions as flexible nonparametric priors over the unknown distributions.

## 1 Introduction

In this paper, we tackle the so-called two-sample problem:

**Problem Statement 1** *Given two samples  $X = \{x_1, \dots, x_{m_1}\} \sim q_1$  and  $Y = \{y_1, \dots, y_{m_2}\} \sim q_2$  from two underlying distributions  $q_1$  and  $q_2$ . The two-sample problem is to decide whether  $q_1 = q_2$ .*

An associated test is called a two-sample test. Such tests are encountered in various disciplines from the life sciences to the social sciences:

- In medical studies, one may want to find out if two classes of patients show different behaviour, response to a drug or susceptibility to a disease.
- In microarray analysis, one may compare measurements from different weeks, labs or platforms to find out if they follow the same distribution, before integrating them into one dataset, in order to increase sample size.
- In the neurosciences, one may want to compare measurements of brain signals under different external stimuli, to check whether brain activity is affected by these stimuli.
- In the social sciences, one may want to compare whether the behavior of a group of people, e.g. when they graduate, marry, or die, is different across countries or generations.
- In the financial sciences, one could for example compare the set of transactions performed at a stock exchange during different weeks, to find out if there is a change in activity in the financial markets.

While this question has been studied in detail by classic statistics for univariate data, there is less work on multivariate data (which we review in Section 2). The only machine learning approach to this problem is a kernel method by (Gretton et al., 2007), using the means of the two samples  $X$  and  $Y$  in a universal reproducing kernel Hilbert Space as its test statistics, but it has created lots of interest in that subject and follow-on studies (Borgwardt et al., 2006; Huang et al., 2007; Gretton & Györfi, 2008).

Here, we approach this two-sample problem from a Bayesian perspective. The classic Bayesian formulation of this problem would be in terms of a Bayes factor (Kass & Raftery, 1995) which represents the likelihood ratio that the data were generated according to hypothesis  $\mathcal{H}_0$  (that is from the same distribution) or hypothesis  $\mathcal{H}_1$  (that is from different distributions). However, how to exactly define these two hypotheses is a crucial question, and many answers have been given in the Bayesian literature with hypotheses that are tailored to a specific problem or application domain; one example are the Bayesian t-tests used in microarray data analysis (Baldi & Long, 2001; Fox & Dimmic, 2006). Our goal in this paper is to define two general classes of two-sample tests that represent a precise formulation of the two-sample problem, but are not tailored to a specific application. They are designed to offer an attractive middle ground between the general idea of using Bayes factors and the specialised hypotheses testing problems studied in the literature.

In detail, we define a class of nonparametric Bayesian two sample tests based on Dirichlet process mixture models. The use of Dirichlet process mixtures for flexible nonparametric modelling of general unknown distributions has a long history in Statistics. However, although the two-sample problem depends crucially on testing whether data come from one or two unknown distributions, Bayesian approaches based on nonparametric density models have not been explored to date. Here we propose and explore such a non-parametric method using the classic Dirichlet process mixture. To the best of our knowledge, the only work that is remotely related is that on a Bayesian test for a parametric versus a nonparametric model of the data by Berger and Guglielmi (Berger & Guglielmi, 1998). This addresses a different but related question since it assumes a parametric null hypothesis. We also define a parametric Bayesian two-sample test where the model of the data is a member of the exponential family. This test generalizes the Bayesian t-test by (Baldi & Long, 2001) and (Fox & Dimmic, 2006), who assume that the samples are Gaussian.

This paper is structured as follows. In Section 2 we will review existing approaches to the two-sample problem on multivariate data, and highlight some differences between frequentist and Bayesian hypothesis testing. In Section 3 we outline the common core of our two Bayesian two-sample test, before providing the details on the parametric test in Section 4 and on the non-parametric test in Section 5.

## 2 Multivariate two-sample tests

**Related work in statistics and kernel machines** Our method is a Bayesian approach to a problem that has been studied in classic statistics and kernel machine learning. Here we describe in short the prominent multivariate two-sample tests (see also (Gretton et al., 2007)).

Frequentist two sample tests follow the same principle of classic hypothesis testing: Given the two samples  $X$  and  $Y$ , a test statistic is computed. Then the distribution of this test statistics under the null distribution ( $q_1 = q_2$ ) is determined. If the value of the test statistic falls into the  $1 - \alpha$ -quantile of the null distribution, the null hypothesis  $q_1 = q_2$  is accepted at significance level  $\alpha$ . If its value exceeds the  $1 - \alpha$  quantile, it is rejected at significance level  $\alpha$ . So the outcome of these test depends on the significance level  $\alpha$  which has to be chosen apriori.

Frequentists tests differ mainly in two points: a) the test statistic they employ and b) the way in which they determine the null distribution for this test statistic. The classic *multivariate t-test* (Hotelling, 1951) assumes that both distributions are multivariate Gaussian with unknown identical covariance; *Friedman and Rafsky* (Friedman & Rafsky, 1979; Henze & Penrose, 1999) define test statistics based on spanning trees, namely the number of edges that connect points from  $X$  to  $Y$  in a minimum spanning tree (Wald-Wolfowitz test) and the closeness of points from  $X$  and  $Y$  in a ranking derived from the minimum spanning tree (Kolmogorov-Smirnow test) (Bickel, 1969; Friedman & Rafsky, 1979). *Rosenbaum's* test statistic is the number of pairs containing a data point from  $X$  and  $Y$  in a minimum distance non-bipartite matching over  $X \cup Y$ . *Hall and Tajvidi* (Hall & Tajvidi, 2002) essentially for each data point count its number of nearest neighbours in  $X \cup Y$  that are from the other sample. *Biau and Györfi's* statistic is the distance between Parzen window estimates of the densities (Anderson et al., 1994; Biau & Györfi, 2005). *Gretton et al.* use the distance between the means of  $X$  and  $Y$  in a universal reproducing kernel Hilbert space as their test statistic (Gretton et al., 2007).

**Frequentist versus Bayesian approach** In contrast to classic hypothesis testing, the test statistic in Bayesian hypothesis testing is a so-called Bayes factor. It is the ratio of the likelihoods of two opposing hypotheses having generated the data  $D = \{X, Y\}$ , the hypothesis  $\mathcal{H}_0$  ( $q_1 = q_2$ ) and its alternative  $\mathcal{H}_1$  ( $q_1 \neq q_2$ ).

To summarize, frequentist classic hypothesis testing considers only one hypothesis and evidence *against* it, whereas Bayesian hypothesis testing compares the likelihoods of two alternative hypotheses having generated the data at hand. While the question of which perspective is to prefer is still an ongoing and unresolved debate, we deem it useful to have a Bayesian alternative to the classic frequentist two sample tests for the following reasons: Bayesian approaches have a clear interpretability compared to the commonly used p-values. Prior knowledge on the probability of the two hypotheses can be incorporated into the Bayes factor in a straightforward manner.

## 3 Concept of Bayesian two-sample tests

### 3.1 Bayes factor as test criterion

Our two classes of Bayesian two sample tests are based on the idea to compute a Bayes factor between two alternative hypotheses: the hypothesis  $\mathcal{H}_1$  that both samples were independently generated from different underlying distributions  $q_1$  and  $q_2$  with  $q_1 \neq q_2$ , and the hypothesis  $\mathcal{H}_0$  that they originated from the same distribution  $q$  ( $q_1 = q_2$ ). This idea is formalised in the following lemma.

**Lemma 1** *Given two samples  $X \sim q_1$  and  $Y \sim q_2$ , we accept the hypothesis  $\mathcal{H}_1$  that  $q_1 \neq q_2$  if the Bayes factor*

$$\chi = \frac{P(X, Y | \mathcal{H}_1)}{P(X, Y | \mathcal{H}_0)} > 1, \quad (1)$$

*otherwise we accept the hypothesis  $\mathcal{H}_0$  that  $q_1 = q_2 = q$ .*

The hypothesis  $\mathcal{H}_1$  is that the samples originate from different distributions, such that

$$P(X, Y | \mathcal{H}_1) = P(X | \mathcal{H}_1)P(Y | \mathcal{H}_1). \quad (2)$$

### 3.2 Computation of the Bayes factor

The central challenge when computing the Bayes factor  $\chi$  is that we do not know the distributions  $q_1$ ,  $q_2$ , and  $q$  our Bayes factor is based upon. Since  $q$ ,  $q_1$  and  $q_2$  are unknown probability distributions, we have to compute the integral over all such distributions with respect to some prior on distributions. We offer two classes of solutions here. In Section 4, we present a parametric test where the distributions are in the exponential family and have conjugate priors. In Section 5, we present a non-parametric test where the distributions  $q_1, q_2$ , and  $q$  are assumed to be drawn from a Dirichlet Process mixture model.

## 4 Parametric Bayesian two-sample test

### 4.1 Exponential Families

For the parametric Bayesian two-sample test, we assume that the underlying distributions  $q_1$  and  $q_2$  are in the exponential family: The distribution for models from this family can be written in the form

$$p(\mathbf{x} | \theta) = f(\mathbf{x})g(\theta) \exp\{\theta^\top \mathbf{u}(\mathbf{x})\}, \quad (3)$$

where  $\mathbf{u}(\mathbf{x})$  is a  $K$ -dimensional vector of sufficient statistics,  $\theta$  are the natural parameters, and  $f$  and  $g$  are non-negative functions. The conjugate prior is

$$p(\theta | \eta, \nu) = h(\eta, \nu)g(\theta)^\eta \exp\{\theta^\top \nu\}, \quad (4)$$

where  $\eta$  and  $\nu$  are hyperparameters, and  $h$  normalizes the distribution.

## 4.2 Bayes factor of parametric test

The Bayes factor of the parametric two-sample test can then be computed as

$$\chi = \frac{P(X|\beta)P(Y|\beta)}{P(X, Y|\beta)} = \quad (5)$$

$$= \frac{\int P(X|\theta)P(\theta|\beta)d\theta \int P(Y|\theta)P(\theta|\beta)d\theta}{\int P(X, Y|\theta)P(\theta|\beta)d\theta} = \quad (6)$$

$$= \frac{h(\eta, \nu) h(\eta + m_1 + m_2, \nu + \mathbf{u}(X) + \mathbf{u}(Y))}{h(\eta + m_1, \nu + \mathbf{u}(X)) h(\eta + m_2, \nu + \mathbf{u}(Y))}, \quad (7)$$

where

$$\mathbf{u}(X) = \sum_{i=1}^{m_1} \mathbf{u}(x_i), \mathbf{u}(Y) = \sum_{j=1}^{m_2} \mathbf{u}(y_j), \quad (8)$$

and  $\beta$  is the set of hyperparameters  $\{\eta, \nu\}$  of the prior.

## 5 Nonparametric Bayesian two-sample test

Unlike its parametric counterpart, our nonparametric Bayesian two-sample test does not employ one single model for the data, but rather the limit of infinitely many components of a finite mixture model:  $P(X|\alpha, \beta) = \int P(X|q)P(q|\alpha, \beta)dq$  where  $q$  is an unknown distribution, modelled as an infinite mixture, and  $\alpha$  and  $\beta$  are hyperparameters controlling it.

This can be achieved via a Dirichlet process mixture of members of the exponential family. The Bayes factor for the nonparametric two-sample test equals

$$\chi = \frac{P(X|\alpha, \beta)P(Y|\alpha, \beta)}{P(X, Y|\alpha, \beta)} \quad (9)$$

where  $P(X|\alpha, \beta)$  is the marginal probability of sample  $X$  under a Dirichlet Process Mixture Model (analogous definitions for  $P(X|\alpha, \beta)$  and  $P(X, Y|\alpha, \beta)$ ) with concentration parameter  $\alpha$  and base measure hyperparameter  $\beta$ .

### 5.1 Dirichlet Process Mixture Models

A key component in our nonparametric two sample test is the ability to approximately infer the marginal probability of a set of observations from a Dirichlet Process Mixture Model (DPM). As these DPMs are at the heart of our nonparametric two-sample test, let us review them here (Ferguson, 1973; Antoniak, 1974).

A Dirichlet Process (DP), and also a Dirichlet Process Mixture Model (DPM), is a probability distribution on probability distributions, and DPMs consider the limit of infinitely many components of a finite mixture model. By allowing for

an infinite number of components, we are able to model the complicated distributions that we encounter in real-world applications via DPMS.

Consider a finite mixture model with  $C$  components

$$p(x^{(i)}|\phi) = \sum_{j=1}^C p(x^{(i)}|\theta_j)p(c_i = j|\zeta) \quad (10)$$

where  $c_i \in \{1, \dots, C\}$  is a cluster indicator variable for data point  $i$ ,  $\zeta$  are the parameters of a multinomial distribution with

$$p(c_i = j|\zeta) = \zeta_j, \quad (11)$$

$\theta_j$  are the parameters of the  $j$ th component, and

$$\phi = (\theta_1, \dots, \theta_C, \zeta). \quad (12)$$

Let the parameters of each component have conjugate priors  $p(\theta|\beta)$  as before, and the multinomial parameters also have a conjugate Dirichlet prior

$$p(\zeta|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/C)^C} \prod_{j=1}^C \zeta_j^{\alpha/C-1} \quad (13)$$

Given a data set  $\mathcal{D} = \{x^{(1)}, \dots, x^{(n)}\}$ , the marginal likelihood for this mixture model is

$$p(\mathcal{D}|\alpha, \beta) = \int \left[ \prod_{i=1}^n p(x^{(i)}|\phi) \right] p(\phi|\alpha, \beta) d\phi, \quad (14)$$

where

$$p(\phi|\alpha, \beta) = p(\zeta|\alpha) \prod_{j=1}^C p(\theta_j|\beta). \quad (15)$$

This marginal likelihood can be rewritten as

$$p(\mathcal{D}|\alpha, \beta) = \sum_c p(c|\alpha) p(\mathcal{D}|c, \beta) \quad (16)$$

where  $c = (c_1, \dots, c_n)$  and

$$p(c|\alpha) = \int p(c|\zeta) p(\zeta|\alpha) d\zeta \quad (17)$$

is a standard Dirichlet integral. The quantity (16) is well-defined even in the limit  $C \rightarrow \infty$ . Although the number of possible settings of  $c$  grows as  $C^n$  and therefore diverges as  $C \rightarrow \infty$ , the number of possible ways of partitioning the  $n$  points remains finite (roughly  $O(n^n)$ ). Using  $\mathcal{V}$  to denote the set of all possible partitioning of  $n$  data points, we can re-write (16) as

$$p(\mathcal{D}|\alpha, \beta) = \sum_{v \in \mathcal{V}} p(v|\alpha) p(\mathcal{D}|v, \beta) \quad (18)$$

## 5.2 Approximate inference of marginal probabilities under DPM

While finite, the number of partitions still grows as  $O(n^n)$  with the size  $n$  of the dataset, rendering an exact inference of the marginal probabilities under a DPM intractable even for moderate size datasets (roughly  $n > 10$ ). Hence we have to resort to approximate inference methods for computing these marginals. One choice is Bayesian hierarchical clustering (BHC), a clustering algorithm that can be used for approximate inference of marginal probabilities under a DPM in  $O(n^2)$  (Heller & Ghahramani, 2005).

## 5.3 Bayes factor in nonparametric test

For the nonparametric two-sample test we use a DPM as the distribution on distributions  $q$ ,  $q_1$ ,  $q_2$ . This allows us to integrate out the parameters of the unknown underlying probability distributions  $q$ ,  $q_1$  and  $q_2$  in a Bayesian manner, while employing a flexible model for these distributions.

The Bayes factor  $\chi$  from (1) can then be computed as

$$\chi = \frac{\int P(X|q_1)P(q_1|\alpha, \beta)dq_1 * \int P(Y|q_2)P(q_2|\alpha, \beta)dq_2}{\int P(X, Y|q)P(q|\alpha, \beta)dq} \quad (19)$$

$$= \frac{P(X|\alpha, \beta)P(Y|\alpha, \beta)}{P(X, Y|\alpha, \beta)}, \quad (20)$$

where  $P(X|q_1) = \prod_{i=1}^{m_1} q_1(x_i)$  and  $P(q_1|\alpha, \beta)$  is a Dirichlet process mixture with concentration parameter  $\alpha$  and base measure hyperparameter  $\beta$ . Hence  $P(X, Y|\alpha, \beta)$  denotes the marginal probability that  $X$  and  $Y$  were generated from this DPM with hyperparameters  $\alpha$  and  $\beta$  (analogous for  $P(X|\alpha, \beta)$  and  $P(Y|\alpha, \beta)$ ).

## 6 Discussion and Conclusions

In this paper, we have proposed two classes of Bayesian two sample tests, a parametric test based on distributions from the exponential family, and a non-parametric test based on Dirichlet Process Mixture Models.

An issue of future work will be the runtime of two-sample tests. Frequentist tests are often expensive to compute, as the test statistic often requires at least an runtime of  $O(n^2)$  for  $n$  datapoints and bootstrapping for determining the null distribution. The Bayesian tests avoid this bootstrapping step and there exist various approximations to a Dirichlet process mixture model (Blei & Jordan, 2005; Kurihara et al., 2006; Kurihara et al., 2007), some of which can be computed in less than  $O(n^2)$ . Hence the Bayesian approach might hold the key for efficient two-sample tests, which we will look at in future work.

## References

- Anderson, N., Hall, P., & Titterton, D. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, *50*, 41–54.
- Antoniak, C. (1974). Mixtures of dirichlet process with applications to bayesian nonparametric problems. *Annals of Statistics*, *2*, 1152–1174.
- Baldi, P., & Long, A. D. (2001). A bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics (Oxford, England)*, *17*, 509–19. PMID: 11395427.
- Berger, J., & Guglielmi, A. (1998). *Bayesian testing of a parametric model versus nonparametric alternatives*.
- Biau, G., & Györfi, L. (2005). On the asymptotic properties of a nonparametric  $l_1$ -test statistic of homogeneity. *IEEE Transactions on Information Theory*, *51*, 3965–3973.
- Bickel, P. (1969). A distribution free version of the Smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, *40*, 1–23.
- Blei, D. M., & Jordan, M. I. (2005). Variational inference for dirichlet process mixtures. *Bayesian Analysis*, *1*.
- Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., & Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics (ISMB)*, *22*, e49–e57.
- Ferguson, T. (1973). A bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*, 209–230.
- Fox, R. J., & Dimmic, M. W. (2006). A two-sample bayesian t-test for microarray data. *BMC bioinformatics*, *7*, 126. PMID: 16529652.
- Friedman, J., & Rafsky, L. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, *7*, 697–717.
- Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola, A. (2007). A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press.
- Gretton, A., & Györfi, L. (2008). Nonparametric independence tests: Space partitioning and kernel approaches. *ALT* (pp. 183–198).
- Hall, P., & Tajvidi, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, *89*, 359–374.



- Heller, K. A., & Ghahramani, Z. (2005). Bayesian hierarchical clustering. *ACM International Conference Proceeding Series*, 119, 297–304.
- Henze, N., & Penrose, M. (1999). On the multivariate runs test. *The Annals of Statistics*, 27, 290–298.
- Hotelling, H. (1951). A generalized t test and measure of multivariate dispersion. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 23–41.
- Huang, J., Smola, A., Gretton, A., Borgwardt, K., & Schölkopf, B. (2007). Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Kurihara, K., Welling, M., & Teh, Y. W. (2007). Collapsed variational dirichlet process mixture models. *IJCAI* (pp. 2796–2801).
- Kurihara, K., Welling, M., & Vlassis, N. A. (2006). Accelerated variational dirichlet process mixtures. *NIPS* (pp. 761–768).