MDL MULTIPLE HYPOTHESIS TESTING

Enes Makalic and Daniel Schmidt

Centre for MEGA Epidemiology, The University of Melbourne,
Level 3, 207 Bouverie Street, Carlton VIC 3053, Australia,
{emakalic,dschmidt}@unimelb.edu.au

ABSTRACT

This paper examines the problem of simultaneously testing many independent multiple hypotheses within the minimum encoding framework. We introduce an efficient coding scheme for nominating the accepted hypotheses in addition to compressing the data given these hypotheses. This formulation reveals an interesting connection between multiple hypothesis testing and mixture modelling with the class labels corresponding to the accepted hypotheses in each test. An advantage of the resulting method is that it provides a posterior distribution over the space of tested hypotheses which may be easily integrated into decision theoretic post-testing analysis.

1. INTRODUCTION

Consider the problem of performing m hypothesis tests from m samples of data $\mathbf{Y}^m = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$, where each sample $\mathbf{y}_i \in \mathcal{Y} \subset \mathbb{R}^{n_i} (1 \leq i \leq m)$. It is assumed that there exist $K \geq 2$ candidate hypotheses under consideration for each test. In a standard frequentist approach to hypothesis testing, one has two candidate hypotheses ($K = 2$) deemed the 'null' and alternative hypothesis respectively, and generally proceeds by performing m independent hypothesis tests. In order to determine whether a null hypothesis is rejected, one must also specify a significance level, $\alpha > 0$, which is often taken to be about $\alpha = 0.05$. The m tests yield p-values $\mathbf{p} =(p_1, \ldots, p_m)$ which may be 'corrected' for multiple testing using, say, Bonferonni-type procedures.

In contrast, minimum encoding methods, such as Minimum Message Length (MML) [1] and Minimum Description Length (MDL) [2, 3], treat all candidate hypotheses on an equal footing (that is, specification of the null and alternative hypotheses is not required) and can automatically determine a suitable significance level solely from the observed data. As an example, given m data sets of non-negative integers, it may be of interest to determine whether each data set was generated by a Poisson, geometric or a negative binomial distribution; that is, there are $K = 3$ candidate hypothesis for each test. To date, such methods have largely been applied to single hypothesis testing and nested model selection problems with great success [4]. It would be of interest if the minimum encoding approach could be extended to the problem of testing multiple hypotheses. This paper considers a minimum encoding approach to the the multiple hypothesis testing problem which can be framed as a special case of latent variable inference. In light of this, the evidence for each hypothesis is deemed to be a latent variable and is inferred from the data. An intriguing consequence of this approach is the close connection to the mixture modelling problem.

The minimum encoding approach to inference advocates choosing the hypothesis that most compresses the data as optimal. This has been formalised in the notion of universal models which are a practical approximation to Kolmogorov complexity [5]. A model $\bar{p}(\cdot)$ is universal relative to a set of distributions $p(\cdot|\boldsymbol{\theta})$ indexed by parameter vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$, if for all $\epsilon > 0$ there exists an $n > 0$ such that

$$\max_{\mathbf{y}^n \in \mathcal{Y}^n} \left\{ \frac{1}{n} \log \frac{p(\mathbf{y}^n | \hat{\boldsymbol{\theta}}_{\rm ML}(\mathbf{y}^n))}{\bar{p}(\mathbf{y}^n)} \right\} \leq \epsilon \tag{1}$$

where $\hat{\boldsymbol{\theta}}_{\rm ML}(\mathbf{y}^n)$ is the maximum likelihood estimator of $\boldsymbol{\theta}$ and $\mathcal{Y}^n \subset \mathbb{R}^n$ is the data space. This implies that the relative difference between the optimal non-transmittable code, $-\log p(\mathbf{y}^n | \hat{\boldsymbol{\theta}}_{\rm ML}(\mathbf{y}^n))$, and the universal model approaches zero as $n \to \infty$; this difference in codelengths is often referred to as *regret* in the literature.

There exist a range of universal models in the literature, including the Normalized Maximum Likelihood [2, 6] code, the Minimum Message Length [1] codes and sequential codes [7]. Under suitable regularity conditions, the codelength for data $\mathbf{y}^n$ using a universal model $\bar{p}(\cdot)$ satisfies

$$-\log \bar{p}(\mathbf{y}^n) = -\log p(\mathbf{y}^n | \hat{\theta}_{\rm ML}(\mathbf{y}^n)) + \frac{k}{2} \log n + O(1) \tag{2}$$

as $n \to \infty$, where $p > 0$ is the number of free parameters. This is the well known Bayesian Information Criterion (BIC) [8]. For finite n, the $O(1)$ term can be arbitrarily large and can have significant effect on inference. Thus, the choice of universal model largely determines the $O(1)$ term for the model class under consideration.

When applying minimum encoding procedures to a single test consisting of K competing hypotheses, one generally proceeds by determining the codelengths of the K candidate models, say $I_k(\mathbf{y}^n)$, and selecting the model with the shortest codelength as the best explanation of the data. Strictly, one also needs a preamble code stating which of the K models, say model k, is subsequently used to compress the data. Let $I(k)$ denote the length of the preamble code. The particular form of a preamble code induces a prior distribution over the support $\{1,\ldots,K\}$. When the number of competing hypotheses K is small, or the models form a nested structure, using an uninformative uniform 'prior' distribution over the K competing hypotheses (that is, $I(k) = \log K$) generally yields satisfactory results [4]. Model selection is then performed by finding the candidate model $\hat{k}$ such that

$$\hat{k} = \arg\min_{k} \{I(k) + I_k(\mathbf{y}^n)\} \tag{3}$$

In words, the 'accepted' hypothesis is the model whose total codelength, which comprises the preamble code, $I(k)$, and the data code $I_k(\mathbf{y}^n)$, is the shortest. Here, the codelengths $I_k(\mathbf{y}^n)$ are assumed to be found by using any suitable universal model.

2. MULTIPLE HYPOTHESIS TESTING

A generalisation of the minimum encoding approach to testing m independent hypotheses is now discussed. Let $\hat{\mathbf{k}}^m \in \{1,\ldots,K\}^m$ denote the set of accepted hypotheses

$$\hat{\mathbf{k}^m} = \arg\min_{\mathbf{k}^m} \left\{ \sum_{i=1}^{m} I(k_i) + I_{k_i}(\mathbf{y}_i^n) \right\}. \tag{4}$$

Choosing a uniform preamble code for each $\hat{k}_i$ (that is, $I(\hat{k}_i) = \log K, 1 \leq i \leq m$), reduces to the single hypothesis testing procedure. While this choice of code expresses prior ignorance about which hypotheses are likely to be 'true', the resulting codelength is optimal only in the case that all K hypotheses are equally likely to occur. In practice, we may expect that one hypothesis is more likely (for example, the conventional 'null' hypothesis) rendering the uniform prior code inefficient in this setting.

We conjecture that a suitable preamble code for $\mathbf{k}^m$ must: (1) attain shorter codelengths than the uniform prior for the majority of data $\mathbf{k}^m \in \{1,\ldots,K\}^m$, (2) be invariant to relabelling of the candidate hypotheses, and (3) be invariant to permutations of the set $\mathbf{k}$. Encoding the accepted hypotheses $\mathbf{k}^m$ as data arising from a multinomial distribution with cell probabilities $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_K)$ satisfies all three requirements. In practice, the ideal cell probabilities are unknown and one may use a suitable universal model to compress the set $\mathbf{k}^m$, obtaining a codelength $I(\mathbf{k}^m)$ which is for almost all cases shorter than the naïve codelength $m \log K$. Even in the worst case of maximum entropy (all $\boldsymbol{\theta}$ being identical), for large m, by use of (2), the codelength $I(\mathbf{k}^m)$ exceeds that of the naïve codelength only by approximately $K/2 \log m$ nits.

To examine the behaviour of this multinomial prior we consider the case of two competing hypotheses (that is, $K = 2$). For m sufficiently large, the codelength for $\mathbf{k}$ is

$$I_m(\mathbf{k}) = -h_1 \log\left(\frac{h_1}{m}\right) - h_2 \log\left(\frac{h_2}{m}\right) + O(\log m) \tag{5}$$

where $h_j = \sum_{i=1}^{m} \mathcal{I}(k_i = j)$ is the number of times hypothesis j is chosen, and $\mathcal{I}(\cdot)$ is the indicator function. The codelength (5) is a symmetric concave function with two minima at $(h_1 = 0, h_2 = m)$ and $(h_1 = m, h_2 = 0)$, and a maximum at $h_1 = h_2$. Some new light may be shed on the behaviour of this prior by instead viewing the issue of variable selection in regression models as a problem of multiple hypothesis testing. Our requirement for invariance under relabelling implies that all hypotheses should be treated on the same footing in terms of a codelength for $\mathbf{k}^m$; thus, there should be no difference between a small number of included regressors or a small number of omitted regressors. Of course, the inclusion of more regressors increases the complexity of the resulting model, but this is captured in the subsequent $\sum_{i=1}^{m} I_k(\cdot)$ codelengths.

A further argument for the merits of the multinomial coding can be drawn from the theory of algorithmic complexity. A so-called 'universal' prior, based on algorithmic complexity, can be defined over a set of strings as $\pi^*(\mathbf{k}^m) \propto \exp(-\mathcal{K}(\mathbf{k}^m|M))$, where $\mathcal{K}(\mathbf{k}^m|M)$ is the algorithmic complexity of the data $\mathbf{k}^m$ with respect to a computer M ([1], pp. 133–135). Such a universal prior is known to have strong theoretical properties; see for example, [9]. Enforcing the invariance restrictions results in a reduced set of possible programs executable by the machine M with which to compress the data $\mathbf{k}^m$, so that the algorithmic complexity is approximately proportional to the Shannon entropy of the data. The set of possible programs is restricted to include only those that assign the same codelength to all permutations of $\mathbf{k}^m$. Thus, the chosen multinomial prior over the set of class allocations exhibits similar behaviour to the universal prior $\pi^*(\cdot)$.

3. MIXTURE MODELLING OF HYPOTHESES

The total codelength $I_h(\mathbf{Y}^m, \mathbf{k})$ for all the datasets $\mathbf{Y}^m = (\mathbf{y}_1,\ldots,\mathbf{y}_m)$ and the chosen hypotheses $\mathbf{k}$ may be written in the following form:

$$\underbrace{-\sum_{i=1}^{m}\sum_{j=1}^{K} \mathcal{I}(k_i = j)\log\theta_j}_{I_h(\mathbf{k})} + \underbrace{\sum_{j=1}^{m}\sum_{j=1}^{K} \mathcal{I}(k_i = j) I_j(\mathbf{y}_i^{n_i})}_{I_h(\mathbf{Y}^m|\mathbf{k})} \tag{6}$$

where $\boldsymbol{\theta}$ denotes the cell probabilities of a K-nomial distribution, and $I_j(\mathbf{y}_i^{n_i})$ is the codelength of dataset $\mathbf{y}_i^{n_i}$ coded using hypothesis j. Strictly, the term $\log 1/\Gamma(K+1)$ should be added to (6) to account for the fact that the labelling of the hypotheses is arbitrary. This encoding is known as the 'hard assignment' codelength in the literature as each dataset is encoded by only one of the K candidate hypothesis.

Examining (6) reveals a close connection with the problem of mixture modelling where the hypothesis chosen to compress each data set is viewed as a latent variable. Thus, for a given a set of competing hypotheses, the multiple hypothesis testing problem is synonymous with estimating class labels of a mixture model in which the K hypotheses take the role of the 'classes', the indicators $\mathbf{k}^m$

are the class labels and the $\mathbf{y}_i^{n_i}$ are the 'data points'. Importantly, treating the data set codelengths as 'likelihoods' and maximising (4) is identical to maximum likelihood estimation of class labels in a regular K-component mixture model which is known to suffer from problems of inconsistency [1]. In multiple hypothesis testing, marginalising the class labels is not an option because the primary reason for doing the testing disappears with the marginalisation.

3.1. Partial Assignment of Hypotheses

The problem of inferring the class labels in a mixture model is synonymous with the allocation of data sets to hypotheses. That is, we can view the class label as a latent variable which determines the optimal hypothesis for a given data set and must be inferred from the available data. The codelengths for $I(\mathbf{k}^m)$ and $I(\mathbf{Y}^m|\mathbf{k}^m)$ are optimal in isolation; this amounts to making an assumption that the $\mathbf{k}^m$ and subsequent compression of the data is independent. However, the choice of $\mathbf{k}^m$ critically depends on the codelengths assigned to the data by the competing hypotheses, so that the assumption of independence results in a codeword longer than could be formulated if the dependence is taken into account.

The reason for the inefficiency is that the hard assignment code assumes the class labels to be known with certainty. However, the class labels are unknown and are themselves being estimated from the data, taking on the role of nuisance parameters. The cardinal rule in the MML and MDL principles is that no parameter should be stated to more accuracy than warranted by the data. While this maxim is easy to interpret in the case of continuous parameters, one should also specify discrete parameters imprecisely. This problem was first identified by Wallace, who subsequently introduced an ingenious way of optimally coding discrete parameters for the problem of regular mixture modelling ([1], pp. 275–295). The new encoding leads to a scheme where both the nuisance class labels as well as the mixture components and their parameters are simultaneously estimated from the data in a consistent fashion.

As an example, consider the task of coding data $\mathbf{y}^n$ with two candidate hypotheses, say $I_1(\cdot)$ and $I_2(\cdot)$. Suppose that the difference in codelength when using hypothesis $I_1(\cdot)$ over the alternative hypothesis is small; that is, $I_1(\mathbf{y}^n) - I_2(\mathbf{y}^n) \leq \epsilon$, where $\epsilon > 0$ is a small. The hard assignment approach accepts hypothesis $I_2(\cdot)$ and ignores the fact that the alternative hypothesis results in a compression that is almost as good. Wallace's 'partial assignment' procedure takes advantage of the fact that the candidate hypotheses yield similar codelengths in encoding the class labels (see [1] for details). Using partial assignment, the total codelength is

$$I_s(\mathbf{Y}^m, \mathbf{k}) = \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{K} r_{ij}\log\frac{r_{ij}}{\theta_j}}_{I_s(\mathbf{k})} + \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{K} r_{ij} I_j(\mathbf{y}_i^n)}_{I_s(\mathbf{Y}^m|\mathbf{k})} \tag{7}$$

where

$$r_{ij} = \frac{\exp\left(-I_j(\mathbf{y}_i^n)\right)\theta_j}{\sum_{q=1}^{K}\exp\left(-I_q(\mathbf{y}_i^n)\right)\theta_q} \tag{8}$$

are the posterior probabilities of data set $\mathbf{y}_i^n$ belonging to hypothesis j for test i; note, as in Section 2, we omit the term $\log 1/\Gamma(K+1)$.

The first term, $I_s(\mathbf{k}^m)$, encodes the class labels (that is, assigns data to hypotheses) optimally based on the posterior probability of the data belonging to the candidate hypotheses. As such, the codelength $I_s(\mathbf{k}^m)$ is always shorter than stating the class labels with absolute certainty as in the hard assignment approach. Conversely, the second term $I_s(\mathbf{Y}^m|\mathbf{k}^m)$ denotes the codelength of data $\mathbf{Y}^m$ under the K different hypotheses using mixture proportions r_{ij}. In hard assignment, since the class label is stated precisely, the codelength of the data, $I_h(\mathbf{Y}^m|\mathbf{k}^m)$, must necessarily be shorter than the corresponding partial assignment code. However, the total codelength $I_s(\mathbf{Y}^m, \mathbf{k}^m)$ is generally significantly shorter than $I_h(\mathbf{Y}^m, \mathbf{k}^m)$, unless the posterior probability of the class labels for each data set is (approximately) one.

Remark 2: Minimising $I_s(\mathbf{Y}^m, \mathbf{k}^m)$ over the class labels $\mathbf{k}^m$, yields the probability of accepting each of the K hypothesis for all m tests. The posterior probability of a hypothesis being accepted can readily be used in a decision theoretic analysis. For example, suppose the tests determine whether a particular drug is to be approved based on m possible side-effects of the drug. The seriousness of each side-effect can be assigned a utility (often of a monetary nature), and based on the posterior probabilities a decision theoretic analysis undertaken to determine whether the drug should be accepted.

3.2. Algorithm

To minimise the partial assignment codelength (7) the following Expectation-Maximisation (EM) [10] type algorithm can be used.

1. Initialise $\boldsymbol{\theta}$ using, for example

$$\theta_j \leftarrow \frac{1}{m}\sum_{i=1}^{m}\mathcal{I}(k_i = \arg\min_q \{I_q(\mathbf{y}^n)\}), \tag{9}$$

   for all $j = 1, 2, \ldots, K$. This is the rate of acceptance of hypothesis j when the effects of multiple testing are not taken into consideration and is equivalent to using a uniform code over the class labels.

2. Update the posterior probabilities using (8) given the mixing proportions $\boldsymbol{\theta}$.

3. Re-estimate $\boldsymbol{\theta}$ by

$$\theta_j \leftarrow \frac{1}{m}\sum_{i=1}^{m} r_{ij}, \ (1 \leq j \leq K). \tag{10}$$

4. Repeat steps (2)–(4) until convergence.

Preliminary empirical testing shows that the algorithm is not particularly sensitive to the initial choice of $\boldsymbol{\theta}$; simulation suggestions that setting $\theta_i = 1/K$ $(1 \leq i \leq K)$ converges to the same solution at roughly the same rate as the initialisation using (9).

3.3. Application Example

A Genome Wide Association Study (GWAS) involves determining whether there exists any genetic association with observable traits; for example, testing whether a particular genetic variant increases the risk of cancer. In the case considered here, one observes a binary data vector $\mathbf{x}^n \in \{0,1\}^n$ that denotes the presence of the trait of interest (also known as the phenotype) in each of the n people, and a matrix $\mathbf{G} \in \{0,1\}^{(n \times m)}$ of measured genetic information (genotypes) for each person. Each of the m columns of the matrix denotes the presence of a genetic mutation at a particular locus in the DNA (known as a single nucleotide polymorphism, or SNP). The aim is to determine whether there exists any association between the m SNPs and the observed data (phenotype) vector $\mathbf{x}^n$.

Under the usual assumption of independence between SNPs, the standard approach is to perform m independent tests of association. For each test, one constructs a 2×2 contingency table from $\mathbf{x}^n$ and $\mathbf{G}$ with entries $\mathbf{y} = \{y_{11}, y_{12}, y_{21}, y_{22}\}$, where the sum of all the entries in a contingency table is equal to n. Given a contingency table, one computes a test statistic, such as χ^2 or Fisher's exact test, and decides that the SNP is associated with the phenotype if this statistic is sufficiently large. An alternative approach based on minimum encoding is to apply the method of Section 3. In a slight shift from the usual way of viewing the problem, we look to test whether the genotype is dependent on the phenotype by compressing the data under two different hypotheses ($K = 2$): (1) the genotype is independent of the phenotype and the data may be compressed concisely by two binomial distributions, and (2) the genotype is dependent on the phenotype, and the data is best compressed using one quadnomial distribution.

Formally, let ϕ_{ij} denote the probability of each cell in a contingency table, such that $\sum_{ij} \phi_{ij} = 1$ for $i, j = 1, 2$, and consider a sampling scheme in which only the sample size n is fixed. Under the independence assumption, $\phi_{ij} = p_i q_j$ where $\sum_{i=1}^{2} p_i = \sum_{i=1}^{2} q_j = 1$ and the contingency table can be compressed using the two parameter universal model for the (constrained) multinomial distribution

$$\binom{n}{\mathbf{y}} p_1^{y_{11}+y_{12}} (1-p_1)^{y_{21}+y_{22}} q_1^{y_{11}+y_{21}} (1-q_1)^{y_{12}+y_{22}}.$$

If the genotype depends on the phenotype vector, the contingency table may instead be compressed using a universal model for the (unconstrained) quadnomial distribution. In both cases, one may use the the Normalised Maximum Likelihood (NML) universal model [2] for which codelengths can be computed in $O(n)$ time using the clever algorithm of [11]; an accurate approximation is given in [12]. The algorithm in Section 3.2 may then be used to minimise the mixture codelength (7) and determine which SNPs are associated with the phenotype. Given the resulting posterior probabilities, one could choose to accept that there is association for SNP j if the corresponding posterior probability is greater than $0{\cdot}5$ $(1 \leq j \leq m)$; otherwise, we accept the hypothesis that the genotype and phenotype for SNP j are independent.

4. REFERENCES

[1] Chris S. Wallace, *Statistical and Inductive Inference by Minimum Message Length*, Information Science and Statistics. Springer, first edition, 2005.

[2] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, January 1996.

[3] Jorma Rissanen, *Information and Complexity in Statistical Modeling*, Information Science and Statistics. Springer, first edition, 2007.

[4] Peter D. Grünwald, *The Minimum Description Length Principle*, Adaptive Communication and Machine Learning. The MIT Press, 2007.

[5] C. S. Wallace and D. L. Dowe, "Minimum message length and Kolmogorov complexity," *Computer Journal*, vol. 42, no. 4, pp. 270–283, 1999.

[6] J. Rissanen, "Strong optimality of the normalized ML models as universal codes and information in data," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1712–1717, July 2001.

[7] Teemu Roos and Jorma Rissanen, "On sequentially normalized maximum likelihood models," in *Proc. 1st Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08)*, Tampere International Center for Signal Processing, 2008, (Invited Paper).

[8] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[9] David L. Donoho, "The Kolmogorov sampler," Tech. Rep. 2002–4, Department of Statistics, Stanford University, 2002.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[11] Petri Kontkanen and Petri Myllymäki, "A linear-time algorithm for computing the multinomial stochastic complexity," *Information Processing Letters*, vol. 103, no. 6, pp. 227–233, September 2007.

[12] W. Szpankowski, *Average case analysis of algorithms on sequences*, John Wiley & Sons, 2001.