A decision-theoretical alternative to testing many hypotheses

NICHOLAS T. LONGFORD

SNTL and Universitat Pompeu Fabra, Ramon Trias Fargas 25–27, 08005 Barcelona, Spain
sntlnick@sntl.co.uk

SUMMARY

Testing a large number of hypotheses is a key problem in the analysis of microarray experiments and in other studies in which many elementary experiments are conducted, and the exceptions among them, for which a particular hypothesis does not hold, have to be identified. A class of approaches to this problem is based on controlling the false discovery rate, even though failure to discover should also be considered. We develop a decision-theoretical approach in which errors (misclassifications) of the two kinds are associated with uneven losses, and the total expected loss in the collection of the classifications (decisions made or options selected) is minimized.

Keywords: Classification; Decision theory; Expected loss; False discovery rate; Hypothesis testing; Penalty ratio.

1. INTRODUCTION

We consider the setting of a large number \( D \) of elementary experiments, called units for short, that comprise a single study. We assume that every unit \( d = 1, \ldots, D \) has an underlying quantity \( \theta_d \), estimated without bias by \( \hat{\theta}_d \), and this estimator is normally distributed with known sampling variance \( s_d^2; \hat{\theta}_d \sim \mathcal{N}(\theta_d, s_d^2) \). The estimators may be correlated. For a large fraction \( p \) of the units, \( \theta_d \) is equal to a known value \( \theta \) or deviates from it by no more than a small quantity \( \delta \), called the largest unimportant deviation. Its value is set by the analyst, but the fraction \( p \) may be unknown. The units for which \( \theta_d \in (\theta - \delta, \theta + \delta) \) are called ordinary, and their complements are the exceptional units. We want to identify all the exceptional units based on the values of \( \hat{\theta}_d \). This is the principal motivating problem for the theory developed by Benjamini and Hochberg (1995) and extended by Benjamini and Yekutieli (2001) and others. See Zhang and Liu (2011) for some recent developments and Moerkerke and others (2006) for an application to DNA marker selection, which raises some issues that we address. The premise of these approaches is that a hypothesis test would be conducted for each unit, and their main result is an adjustment of the rejection regions of these tests by which the false discovery rate is controlled in a prescribed way. A logical inconsistency arises when units for which the hypothesis is not rejected are subsequently treated as if they satisfied the hypothesis, without any evidence to support it.

We want to treat the two kinds of error, false discovery and failure to discover, more evenhandedly by incorporating in the analysis their relative gravity (seriousness). For this purpose, we apply decision theory (Berger, 1985; Lindley, 1985; DeGroot, 2004). For applications to quality control and medical
screening, see Longford (2010) and Longford (2013), respectively. We formulate criteria for the individual classifications. Scott and Berger (2006) and Müller and others (2007) survey methods with criteria centered on the rates of the two kinds of error; see also Wu and Peña (2013). We are concerned solely with finite-sample properties. For asymptotic theory, we refer the reader to Cohen and Sackrowitz (2005) and Bogdan and others (2011).

We argue that the problem of multiple testing disappears when minimization of the expected total loss is pursued, because the operations of expectation, addition, and minimization can be interchanged. However, the loss function has to be universal and additive. That is, the loss of one utile in one instance has to be for all purposes equivalent to the loss of one utile in another instance, and the loss of \( a + b \) utiles in one instance equivalent to the loss of \( a \) utiles in one and \( b \) utiles in another instance. Without satisfying these conditions, neither the total of the losses over a set of units nor their expectation are meaningful quantities. With additivity, each unit is classified by a separate minimization, and the minima are added up to establish the minimum total expected loss. This is simpler than the operations in testing the corresponding unit-related hypotheses, which involve a more difficult non-additive calculus of \( p \) values. The units or their subsets may have different loss functions. The assumption of additivity may become onerous in such a setting, especially with many subsets, and loss functions in a wide range.

Correlation of the estimators \( \hat{\theta}_d \) does not effect the decision for any particular unit, nor the expected total loss, but it has an impact on the variance of the total loss. This may be a concern in the design of a study (Müller and others, 2004), but not in its analysis, especially when units are assessed sequentially. Estimators \( \hat{\theta}_d \) may be improved by exploiting their correlation structure, but we can assume that it has been done already, without introducing any bias.

Potential applications of our approach include assessment of performance of institutions and disease mapping. In contrast to Ohlssen and others (2007), our principal motivation is not a search for outliers. Several values of \( \theta_d \) may be close to and on either side of \( \pm \delta \), especially when \( \delta \) is a standard set prior to the study. Unlike Catelan and others (2010), we regard the loss function as a key input into the analysis (Berry and Hochberg, 1999), and regard a hypothesis test as not useful when a choice between two courses of action is sought.

The next section works out the details in a frequentist perspective, in which \( \delta \) is the only prior information used. In the following section, a Bayesian version of the problem is solved, using a prior distribution for the collection of values of \( \theta_d, d = 1, \ldots, D \). In the example in Section 4, we compare the two versions and study their sensitivity with respect to the prior information. An application to gene expression is presented in Section 5. The paper is concluded by a discussion.

2. Classification

Suppose that a false discovery is associated with the loss of one utile (the unit of loss), and an incorrect omission with \( R \) utiles; \( R \) is referred to as the penalty ratio. The loss function for this setting, called piecewise constant, is defined as \( L(\hat{\theta}_d, \theta_d) = 1 \) for false discovery, as \( L(\hat{\theta}_d, \theta_d) = R \) for incorrect omission, and as \( L(\hat{\theta}_d, \theta_d) = 0 \) otherwise, when the classification is correct. The first argument of \( L \) can be replaced by the decision made, but we assume that it is based solely on \( \hat{\theta}_d \). Our development is for \( R > 0 \), but in a typical application \( R > 1 \) --- an incorrect omission is a more serious error because the unit involved is lost irretrievably, whereas the falseness of a discovery may be established later. The constants \( \delta \) and \( R \) are often difficult to set with integrity, so we solve the problem for ranges of their plausible values. As a result, the classification of a unit may be equivocal --- for some plausible pairs \((\delta, R)\) it is classified as ordinary and for others as exceptional. When there are only a few such units, the uncertainty about \( \delta \) and \( R \) is acceptable.

Prior to realizing the study, \( \hat{\theta}_d \) for a given unit \( d \) is a random variable. After its realization, \( \hat{\theta}_d \) is known. For a frequentist \( \theta_d \) is fixed, and for a Bayesian it is random. Let \( \varepsilon \) be a standard normal variate such that
\( \hat{\theta}_d = \theta_d + s_d \varepsilon \).

Then \( \hat{\theta}_d = \hat{\theta}_d + s_d \varepsilon' \), where \( \varepsilon' = -\varepsilon \), so the distribution of \( \varepsilon' \) is also \( \mathcal{N}(0, 1) \). Irrespective of how we regarded \( \theta_d \) earlier, it is now a random variable. This change of status is natural to a Bayesian (using a flat improper prior), but controversial in the frequentist paradigm; see Seidenfeld (1992) and Hannig (2009) for background and reviews.

Denote by \( \phi(x) \) the density and by \( \Phi(x) \) the distribution function of the standard normal distribution. We use the notation \( \phi(y; \hat{\theta}_d, s_d) = \phi((y - \hat{\theta}_d)/s_d)/s_d \) for the density of \( \mathcal{N}(\hat{\theta}_d, s_d^2) \). If we classify unit \( d \) as exceptional, then the expectation of the loss we incur is

\[
Q_X = \int_{\theta - \delta}^{\theta + \delta} \phi(y; \hat{\theta}_d, s_d) \, dy = \Phi(z_d^+) - \Phi(z_d^-),
\]

where \( z_d^+ = z_d^0 + \delta/s_d \), \( z_d^- = z_d^0 - \delta/s_d \), and \( z_d^0 = (\theta - \hat{\theta}_d)/s_d \), so that \( \phi(\theta; \hat{\theta}_d, s_d) = \phi(z_d^0)/s_d \). The expected loss when classifying a unit as ordinary is \( Q_O = R(1 - Q_X) \). We choose the category for which the expected loss is smaller, that is, depending on the sign of the balance, defined as \( \Delta Q = Q_X - Q_O = (R + 1)Q_X - R \); exceptional if \( \Delta Q < 0 \) and ordinary if \( \Delta Q > 0 \).

We derive an equivalent rule that bypasses the evaluation of \( \Delta Q \) and is better suited for classifying many units. We write \( \Delta Q = \Delta Q(\hat{\theta}_d, s_d) \) to indicate that we are interested in \( \Delta Q \) as a function of the two arguments. It is easy to check that, for any \( s_d \) fixed to a value \( s > 0 \), \( \Delta Q(\hat{\theta}_d, s) \) is symmetric around \( \theta \); it increases up to its maximum at \( \hat{\theta}_d = \theta \), and then decreases. Its limit is \( -R \) as \( \hat{\theta}_d \) diverges to \( +\infty \) or \(-\infty \). Therefore, a unit with \( s_d = s \) is classified as exceptional, irrespective of its value of \( \hat{\theta}_d \), when \( \Delta Q(\theta, s) \) is negative, that is, when

\[
s > s^* = \frac{\delta}{\Phi^{-1}((2R + 1)/(2R + 2))}.
\]

The \textit{borderline} function \( g(s) \) is defined as the positive root of the balance equation \( \Delta Q[g(s) + \theta, s] = 0; s \in [0, s^*) \). The other root is \( -g(s) \). Unit \( d \) is classified as ordinary if \( |\hat{\theta}_d - \theta| < g(s_d) \) and \( s_d < s^* \), and as exceptional otherwise. The values of \( g \) can be found by the Newton–Raphson (NR) or a similar algorithm. Although \( g(s) \) is evaluated iteratively, classification of many (thousands) of units is faster using \( g(s_d) \), especially when \( s_d, d = 1, \ldots, D \), are in a narrow range or attain only a few distinct values. The NR algorithm requires fewer than 10 iterations even for some esoteric values of \( R \) and \( \delta \) and very strict convergence criteria.

The two panels at the top of Figure 1 display the functions \( g(s) \), \( s \leq 0.20 \), for \( \delta = 0.2 \) and \( 0.4 \) and \( R = 1, 2, \ldots, 10 \). Note that \( R \) does not have to be an integer. When \( \theta_d \) is known, \( s = 0 \) and \( g(s) = \delta \), so unit \( d \) is classified straightforwardly. For \( R = 1 \), when errors of the two kinds are equally serious, \( g(s) \approx \delta \) for small \( s \), because the uncertainty about \( \theta_d \) can be ignored. For greater \( \delta \), the curve \( g(s) \) starts to descend for greater \( s \); for \( \delta = 0.4 \) (and \( R = 1 \)), the uncertainty can be ignored even for \( s = 0.30 \).

Every borderline curve \( g(s) \) in Figure 1 are defined at the critical value \( s^* = s^*(\delta, R) \) and \( g(s) \to 0 \) as \( s \to s^* \). In several cases, \( s^* \) is off the horizontal scale. In the top right-hand panel of Figure 1, \( s^*(0.4, R) > 0.2 \) even for \( R = 10 \). In accord with (2.1), \( s^* \) increases in \( \delta \) and decreases in \( R \). Smaller values of \( \delta \) and greater values of \( R \) correspond to more inclusive criteria for exceptionality, and therefore to stronger preference for classifying a unit as exceptional. For sufficiently large \( s_d \), we prefer the classification “exceptional” because \( Q_X = P[\theta_d \in (\theta - \delta, \theta + \delta) | \hat{\theta}_d; s_d] \) is very small.

The values of \( \delta \) and \( R \) are elicited from subject-matter experts; \( \delta \) defines the meaning of “exceptional” and \( R \) quantifies the consequences of an incorrect omission in relation to a false discovery. They should be informed by the role the discoveries play in the subsequent research, business, or operational agenda. Similar issues arise in medical screening (Molanes-López and Letón, 2011; Longford, 2013), to which our problem has obvious affinity.
Fig. 1. The borderline functions $g(s)$ for the piecewise constant, linear, and quadratic loss (rows) with $\delta = 0.2$ and 0.4 (columns) and $R = 1, 2, \ldots, 10$ (indicated in the panels).
The piecewise constant loss is a special case of piecewise power loss functions, defined as \((\delta - |\theta_d - \theta|)^h\) for false discovery, when \(\theta_d \in (\theta - \delta, \theta + \delta)\) but we conclude otherwise, and as \(R(|\theta_d - \theta| - \delta)^h\) for incorrect omission in the converse case. Loss functions with \(h > 0\) are appropriate when not only the kind of the error matters, but also its magnitude,

\[
e_d = \min(|\hat{\theta}_d - \theta - \delta|, |\hat{\theta}_d - \theta + \delta|).
\] (2.2)

Apart from \(h = 0\), only the exponents \(h = 1\) and 2, defining the respective linear and quadratic kernels, are of any practical importance. In what follows, we use the identities \(\int z\phi(z) \, dz = -\phi(z)\) and \(\int \Phi(z) \, dz = z\Phi(z) + \phi(z)\). The latter function is denoted by \(F_1\). With the linear kernel, the expected loss due to false discovery is

\[
Q_X = \int_{\theta-\delta}^{\theta+\delta} (\delta + \theta - y)\phi(y; \hat{\theta}_d, s_d) \, dy - \int_{\theta-\delta}^{\theta+\delta} (\delta - \theta + y)\phi(y; \hat{\theta}_d, s_d) \, dy
\]

\[
= s_d\{F_1(z^{+}_d) + F_1(z^{-}_d) - 2F_1(z^0_d)\},
\]

obtained by integrating by parts. The expected loss when classifying a unit as ordinary is

\[
Q_O = R \left\{ \int_{\theta+\delta}^{+\infty} (y - \theta - \delta)\phi(y; \hat{\theta}_d, s_d) \, dy + \int_{-\infty}^{\theta-\delta} (\theta - \delta - y)\phi(y; \hat{\theta}_d, s_d) \, dy \right\}
\]

\[
= R s_d\{-z^{+}_d + F_1(z^{+}_d) + F_1(z^{-}_d)\}.
\]

We seek the roots of the balance function

\[
\Delta Q = Q_X - Q_O = s_d\{Rz^{+}_d - (R - 1)\{F_1(z^{+}_d) + F_1(z^{-}_d)\} - 2F_1(z^0_d)\},
\] (2.3)

which we use for classifying the units. The NR algorithm can be applied with advantage, because

\[
\frac{\partial \Delta Q}{\partial \hat{\theta}_d} = (R - 1)\{\Phi(z^{+}_d) + \Phi(z^{-}_d)\} + 2\Phi(z^0_d) - R
\]

is a simple expression. For a value of \(s_d\) fixed at \(s\), the function \(\Delta Q(\hat{\theta}_d, s)\) is symmetric around \(\hat{\theta}_d = \theta\), and it increases for \(\hat{\theta}_d < \theta\) and decreases for \(\hat{\theta}_d > \theta\). Therefore, if \(\Delta Q(\theta, s) > 0\), \(\Delta Q\) has a single pair of roots \(\theta \pm g(s)\), and no roots otherwise. When the roots exist, we classify unit \(d\) with \(s_d = s\) as ordinary if \(|\hat{\theta}_d - \theta| < g(s)\), and as exceptional otherwise.

To explore when \(\Delta Q\) has no roots, let \(t = \delta/s\). When \(\hat{\theta}_d = \theta\) and \(s_d = s\), \(z^0_d = 0\), and \(z^{\pm}_d = \pm t\). The function \(q(t) = \Delta Q(\theta, s)/s = Rt - (R - 1)\{2F_1(t) - t\} - 2\phi(0)\) is increasing for all \(t > 0\), is negative at \(t = 0\), and it diverges to \(+\infty\) as \(t \to +\infty\). Therefore, \(q(t)\) has a unique root for every \(\delta > 0\) and \(R > 0\), denoted by \(t^*(q, R)\). For \(t < t^*(q, R)\), or equivalently, for \(s > s^* = \delta/t^*(q, R)\), the function \(\Delta Q(g, s)\) has no roots for \(g\). A unit with \(s_d > s^*\) is classified as exceptional irrespective of its value \(\hat{\theta}_d\). The critical value \(s^*\) can be found by the NR algorithm.

The middle row of Figure 1 displays the borderline functions \(g(s)\) for penalty ratios \(R = 1, \ldots, 10\) and largest unimportant differences \(\delta = 0.2\) and 0.4. If \(|\hat{\theta}_d - \theta| > g(s_d)\), or \(g(s_d)\) is not defined because \(s_d > s^*\), the unit is classified as exceptional; \(s^*\) depends on \(\delta\) and \(R\).
For quadratic kernel, the expected losses due to the two kinds of error are

\[
Q_X = \int_{\theta}^{\theta+\delta} (\theta + \delta - y)^2 \phi(y; \hat{\theta}_d, s_d) \, dy + \int_{\theta-\delta}^{\theta} (y - \theta + \delta)^2 \phi(y; \hat{\theta}_d, s_d) \, dy
\]

\[
= s_d^2 \left\{ \int_{z_d^+}^{z_d^+} (z_d^+ - z)^2 \phi(z) \, dz + \int_{z_d^-}^{z_d^-} (z - z_d^-)^2 \phi(z) \, dz \right\}
\]

\[
= s_d^2 \left\{ F_2(z_d^+) - F_2(z_d^-) - \frac{4\delta}{s_d} F_1(z_d^0) \right\},
\]

\[
Q_O = R \int_{\theta+\delta}^{\infty} (y - \theta - \delta)^2 \phi(y; \hat{\theta}_d, s_d) \, dy + R \int_{-\infty}^{\theta-\delta} (\theta - y - \delta)^2 \phi(y; \hat{\theta}_d, s_d) \, dy
\]

\[
= Rs_d^2 \{ 1 + z_d^+^2 - F_2(z_d^+) + F_2(z_d^-) \},
\]

where \( F_2(z) = (1 + z^2) \Phi(z) + z \phi(z) = 2 \int F_1(z) \, dz \). The balance function \( Q_X - Q_O \) is

\[
\Delta Q = s_d^2 \left[ (R + 1)[F_2(z_d^+) - F_2(z_d^-)] - R(1 + z_d^+^2) - \frac{4\delta}{s_d} F_1(z_d^0) \right]. \tag{2.4}
\]

This function of \( \hat{\theta}_d \) (involved in \( z_d^+ \)) is symmetric around \( \theta \). Its derivatives with respect to \( \hat{\theta}_d \) are

\[
\frac{\partial \Delta Q}{\partial \hat{\theta}_d} = 2s_d \left[ Rz_d^+ - (R + 1)[F_1(z_d^+) - F_1(z_d^-)] + \frac{2\delta}{s_d} \Phi(z_d^0) \right],
\]

\[
\frac{\partial^2 \Delta Q}{\partial \hat{\theta}_d^2} = 2(R + 1)[\Phi(z_d^+) - \Phi(z_d^-)] - 2R - \frac{4\delta}{s_d} \phi(z_d^0). \tag{2.5}
\]

The first-order derivative is an odd function of \( \hat{\theta}_d - \theta \). It vanishes at \( \hat{\theta}_d = \theta \), when \( z_d^0 = 0 \), and diverges to \( +\infty \) for \( \hat{\theta}_d \to \pm \infty \). However, \( \partial \Delta Q / \partial \hat{\theta}_d \) has also other roots for some configurations of \( t = \delta/s_d \) and \( R \), for which the second-order derivative is not negative throughout. In such cases, classification based on the borderline function \( g(s) \) cannot be applied and \( \Delta Q \) has to be evaluated separately for every unit. Details and some insight into how this happens are relegated to supplementary material available at Biostatistics online (http://www.biostatistics.oxfordjournals.org). For \( R \in (1, 10) \) and \( \delta = 0.2 \) and 0.4, this problem does not arise, and the corresponding functions \( g(s) \) are displayed in the bottom row of Figure 1.

A key feature of the loss functions we use is asymmetry \( (R \neq 1) \). The variety of loss structures can be extended by using different kernels for the two kinds of error, such as the constant loss for false discovery and linear or quadratic loss for incorrect omission. The penalty ratio may also be set to different values for subsets of the units. Further, two loss functions, say, \( L_1 \) and \( L_2 \), can be combined. For any positive constants \( a_1 \) and \( a_2 \), \( a_1 L_1 + a_2 L_2 \) is also a loss function. The calculus for comparing the expected losses is easy to adapt for such composite loss functions, because the operations of integration (or differentiation) and linear combination can be interchanged.

The piecewise power loss functions are distantly related to hypothesis testing (absolute kernel), estimation with the mean squared error (quadratic kernel), and estimation with the mean absolute error (linear kernel), and so they are easy to motivate for a client (expert) whose input is essential for specifying the kernel, the penalty ratio \( R \), and the largest unimportant deviation \( \delta \). The outcome of elicitation from a client may be a range of plausible values of \( \delta \) and \( R \), say, \( (\delta_L, \delta_U) \) and \( (R_L, R_U) \). In that case, we solve the problem for the limits of these ranges, the four vertices of the plausible rectangle \( (\delta_L, \delta_U) \times (R_L, R_U) \), and classify a unit as ordinary or exceptional if it is classified so at every vertex. Otherwise, we reach an
impasse. If it occurs for a non-trivial fraction of the units, then the elicitation should be revisited to narrow down the plausible ranges.

The absolute kernel loss function has a discontinuity, the linear kernel loss is not smooth and the derivative of the quadratic kernel is not smooth at zero. This raises no analytical difficulties in the classification. An example of a class of smooth loss functions is the LINEX (Zellner, 1986) defined as $L_a(x) = \exp(ax) - ax - 1$, where the asymmetry coefficient $a \neq 0$ is a parameter, and the error $x$ is defined below. For $a > 0$, $L_a(x) \gg L_a(-x)$ for large positive $x$. Therefore, we specify $x$ so that it is positive for incorrect omission, $x = e_d$ (see (2.2)) and as $x = -e_d$ for false discovery. The expected loss with $L_a(x)$ is derived in supplementary material available at Biostatistics online (http://www.biostatistics.oxfordjournals.org). The result is a rather unwieldy expression. A drawback of LINEX is that the description of its asymmetry is not as simple as for the power kernels, for which the ratio of losses $L(e_d)/L(-e_d) = R$ is constant.

3. PRIOR INFORMATION

Although $\delta$ constitutes some prior information, it does not fall into the framework of a Bayesian analysis. In this section, we consider prior information about the ordinary and exceptional units, expressed in terms of their prior probabilities and distributions.

Suppose that the values $\theta_d$ for the ordinary units have a normal prior distribution with density $f_O$ and the exceptional units a mixture of two normal distributions with densities $f_L$ and $f_U$. We assume first that the probabilities of these three components, $p_O$, $p_L$, and $p_U$, are known. The two components of the mixture for the exceptional units may be mirror images around $\theta$: $p_L = p_U = \frac{1}{2} p_X$, and $\mathcal{N}(\theta + \gamma, \sigma^2_X)$ and $\mathcal{N}(\theta - \gamma, \sigma^2_X)$ for some $\gamma > \delta$. We proceed as follows. From the (conditional) distribution of $(\hat{\theta}_d | \theta_d)$, $\mathcal{N}(\theta_d, s_d^2)$, we derive the posterior distribution of $(\theta_d | \hat{\theta}_d)$, which turns out to be a mixture of normals, and evaluate the expected loss with respect to this mixture. Then we choose the option for which the posterior expected loss is smaller.

For notational simplicity, and to derive a more general result, suppose that the distribution of all the units is a mixture of $K$ normal distributions $\mathcal{N}(\mu_k, \sigma_k^2)$, with respective marginal probabilities $p_k, k = 1, \ldots, K$. The posterior density of $\theta_d$ for unit $d$ is

$$
\xi(\theta_d; \hat{\theta}_d) = C \phi(\hat{\theta}_d; \theta_d, s_d^2) \sum_{k=1}^K p_k \phi(\theta_d; \mu_k, \sigma_k^2),
$$

where $C$ is the normalizing constant. By consolidating the arguments of the exponentials in the normal densities, we obtain the expression

$$
\xi(\theta_d; \hat{\theta}_d) = \sum_{k=1}^K \bar{p}_k \phi(\hat{\theta}_d; \mu_k, \sigma_k^2),
$$

where

$$
\bar{p}_k = \frac{p_k \phi(\hat{\theta}_d; \mu_k, \sigma_k^2)}{\sum_{h=1}^K p_h \phi(\hat{\theta}_d; \mu_h, \sigma_h^2)},
$$

$$
\tilde{\mu}_k = \frac{\hat{\theta}_d \sigma_k^2 + \mu_k s_d^2}{\sigma_k^2 + s_d^2}, \quad \tilde{\sigma}_k^2 = \frac{s_d^2 \sigma_k^2}{s_d^2 + \sigma_k^2}, \quad \text{and} \quad \bar{p}_k = \frac{p_k \phi(\hat{\theta}_d; \mu_k, \sigma_k^2)}{\sum_{h=1}^K p_h \phi(\hat{\theta}_d; \mu_h, \sigma_h^2)}.
$$

Thus, the posterior distribution of $\theta_d$ is also a mixture of $K$ normals, but both the component distributions and the probabilities depend on $\hat{\theta}_d$ and $s_d$.

Denote the posterior probabilities of the three components for unit $d$ by $\bar{p}_{dO}$, $\bar{p}_{dL}$ and $\bar{p}_{dU}$ for being an ordinary unit and an exceptional unit in the left-hand and right-hand component, respectively.
posterior means and variances, we use the notation \( \tilde{\theta}_{dt} \) and \( \tilde{\sigma}_{dt}^2 \) with T = O, L, or U. The posterior expected loss for false discovery of unit d with the quadratic kernel is

\[
Q_{dX} = \tilde{p}_{dO} \int_{\theta}^{\theta+\delta} (\theta + \delta - y)^2 \phi(y; \tilde{\mu}_{dO}, \tilde{\sigma}_{dO}^2) dy + \tilde{p}_{dO} \int_{\theta-\delta}^{\theta} (y - \theta + \delta)^2 \phi(y; \tilde{\mu}_{dO}, \tilde{\sigma}_{dO}^2) dy
\]

\[
= \tilde{p}_{dO} \tilde{\sigma}_{dO}^2 \left\{ F_2(u_{dO}^+) - F_2(u_{dO}^-) - \frac{4\delta}{\tilde{\sigma}_{dO}} F_1(u_{dO}) \right\},
\]

where \( u_{dO}^0 = (\theta - \tilde{\mu}_{dO})/\tilde{\sigma}_{dO} \) and \( u_{dO}^\pm = u_{dO}^0 \pm \delta/\tilde{\sigma}_{dO} \) are the Bayesian versions of the quantities \( z_d^0 \) and \( z_d^\pm \) defined in the previous section, except that now they are specific for the “ordinary” component and unit d. The quantities \( u_{dL}^\pm \) and \( u_{dU}^\pm \) are defined similarly. The expressions for \( Q_{dL} \) and \( Q_{dU} \) are obtained by applying the same calculus as in Section 2:

\[
Q_{dU} = \tilde{p}_{dU} R \int_{\theta+\delta}^{+\infty} (y - \theta - \delta)^2 \phi(y; \tilde{\mu}_{dU}, \tilde{\sigma}_{dU}) = \tilde{p}_{dU} R \tilde{\sigma}_{dU}^2 (1 + u_{dU}^+ - F_2(u_{dU}^-)),
\]

\[
Q_{dL} = \tilde{p}_{dL} R \int_{-\infty}^{\theta-\delta} (\theta - \delta - y)^2 \phi(y; \tilde{\mu}_{dL}, \tilde{\sigma}_{dL}) = \tilde{p}_{dL} R \tilde{\sigma}_{dL}^2 F_2(u_{dL}^-).
\]

The posterior expected loss for the incorrect omission of unit d is \( Q_{dO} = Q_{dL} + Q_{dU} \). For the posterior balance \( Q_{dX} - Q_{dO} \), the approach based on the borderline functions cannot be applied; the expected losses \( Q_{dX} \) and \( Q_{dO} \) have to be compared directly.

For the linear kernel, we have the identities

\[
Q_{dX} = \tilde{p}_{dO} \tilde{\sigma}_{dO} [F_1(u_{dO}^+) + F_1(u_{dO}^-) - 2F_1(u_{dO}^0)],
\]

\[
Q_{dO} = R [\tilde{p}_{dL} \tilde{\sigma}_{dL} F_1(u_{dL}) - \tilde{p}_{dU} \tilde{\sigma}_{dU} u_{dU} - F_1(u_{dU}^+)],
\]

and for the absolute kernel

\[
Q_{dX} = \tilde{p}_{dO} (\Phi(u_{dO}^+) - \Phi(u_{dO}^-)),
\]

\[
Q_{dO} = R [\tilde{p}_{dL} \Phi(u_{dL}) + \tilde{p}_{dU} (1 - \Phi(u_{dU}))].
\]

The prior as well as the posterior distributions may be in conflict with the assumption that units with \( \theta_d \in (\theta - \delta, \theta + \delta) \) are ordinary, because their supports are not bounded. In any reasonable setting, the probability of this occurring is small. In the derivations of \( Q_{dX} \) and \( Q_{dO} \), we assumed that such contradictions do not arise. In practice, if they did arise, we would regard a unit in component O, for which \( \theta_d \notin (\theta - \delta, \theta + \delta) \), as exceptional. A prior can be specified for the marginal probabilities \( p_O \), \( p_L \), and \( p_U \); it may be degenerate if \( p_L = p_U \) is assumed. An alternative to such a prior is a sensitivity analysis with a range of plausible values of the probabilities.

4. Example

We generated a dataset of \( D = 5000 \) units by the following scheme. The ordinary units comprise 3900 units with \( \theta_d = \theta = 0 \) and 1000 units with values of \( \theta_d \) generated as a random sample from \( N(0, 0.01) \). The values of \( \theta_d \) for the 100 exceptional units are drawn from the mixture of \( N(-1.0, 0.04) \) and \( N(1.0, 0.0625) \), with respective probabilities 0.70 and 0.30. We set first \( \delta = 0.40 \). The value for one exceptional unit had to be adjusted to be outside the range \((-0.4, 0.4) \). The sampling variances \( \hat{s}_{d}^2 \) of the estimators \( \hat{\theta}_{dt} \) are 0.0064 for the ordinary and 0.0100 for the exceptional units. We simulate their estimates \( \hat{\theta}_{d}^2 \) by random draws.
from scaled $\chi^2$ distributions, with the degrees of freedom (d.f.) themselves drawn at random from the uniform distribution on (25, 40), independently from the draws from $\chi^2$. The generated dataset is displayed in Figure 2(A). The 1100 non-zero estimates $\hat{\theta}_d$ are indicated by black dots, at height 0.4 for the 1000 ordinary units and at 0.6 for the 100 exceptional units. Uniformly distributed vertical noise is added to the values of the ordinary units to avoid massive overprinting. The underlying values $\theta_d$ are marked by gray dots at the bottom and top margin for the ordinary and exceptional units, respectively. The 3900 units with $\theta_d = 0$ and their values $\hat{\theta}_d$ are omitted from the diagram. Each unit is associated with a vertical gray segment of length proportional to $\hat{\sigma}_d$, centered at the coordinates of the estimate $\hat{\theta}_d$. The values of $\theta_d$ and $\hat{\theta}_d$ for the exceptional units are connected by thin segments at height 0.70–0.95, to indicate the estimation errors $\hat{\theta}_d - \theta_d$. The values of $\hat{\delta}_d$ for the 5000 units are in the range (0.048, 0.135). These are direct estimates, with no pooling of information across the units. Their mean and median are both equal to 0.080.

Figure 2(B) displays the borderline functions $g(s)$ in the range of the realized values of $\hat{\delta}_d$ for the linear kernel with $R = 10$ and 20, and $\delta = 0.4$ and 0.5. Many of the values $|\hat{\theta}_d|$ are outside the plotting range, but the exceptional unit with the smallest value of $|\hat{\theta}_d|$, equal to 0.346, is within the range. The three units marked by crosses on gray background have $0.4 < |\theta_d| < 0.5$, so they are reclassified from exceptional to
Table 1. The losses for false discovery ($L_X$), incorrect omission ($L_O$) and their totals ($L$) by the frequentist (F) and two Bayesian methods (B1 and B2) in a single replication and their averages ($Q_X$, $Q_O$, and $Q$) over 500 replications

<table>
<thead>
<tr>
<th>Method</th>
<th>Loss in one replication</th>
<th>Empirical expected loss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\delta$</td>
<td>$R$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>$L_X$</td>
<td>2.6815</td>
</tr>
<tr>
<td></td>
<td>$L_O$</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>$L$</td>
<td>2.6815</td>
</tr>
<tr>
<td>B1</td>
<td>$L_X$</td>
<td>0.4020</td>
</tr>
<tr>
<td></td>
<td>$L_O$</td>
<td>1.2620</td>
</tr>
<tr>
<td></td>
<td>$L$</td>
<td>1.6640</td>
</tr>
<tr>
<td>B2</td>
<td>$L_X$</td>
<td>0.4020</td>
</tr>
<tr>
<td></td>
<td>$L_O$</td>
<td>1.2620</td>
</tr>
<tr>
<td></td>
<td>$L$</td>
<td>1.6640</td>
</tr>
</tbody>
</table>

The subscripts indicate the number of misclassified units; the numbers of these units that were reclassified, for which $0.4 < |\hat{\theta}_d - \theta| < 0.5$, are given in parentheses.

ordinary as $\delta$ is raised from 0.4 to 0.5. Note that their values of $|\hat{\theta}_d|$ are not the three smallest among the exceptional units.

The borderline functions $g(s)$ for $(\delta, R) = (0.4, 20)$ and $(0.5, 10)$ delimit the gray zone within which the classification is equivocal. One exceptional, one reclassified and 20 ordinary units are in the gray zone. The appropriate option is chosen for the reclassified unit in all four instances, because its value of $|\hat{\theta}_d|$, 0.346, is above the borderlines for $\delta = 0.4$ (and both $R = 10$ and 20), when it is an exceptional unit, and is under the borderlines for $\delta = 0.5$ when it is an ordinary unit. Two ordinary units are misclassified for all plausible pairs $(\delta, R)$.

The losses are listed in the left-hand part of the first block of Table 1 (Method F). For false discovery when $\delta = 0.4$, they are $L = L_X + L_O = 2.68$ and $L = 4.50$ for $R = 10$ and 20, owing to 15 and 22 misclassified units, respectively. There is no incorrect omission in either setting. For $\delta = 0.5$, the loss for false discovery is 0.70 for both $R = 10$ and 20, owing to four units, two of them reclassified. For $R = 10$, there is an additional loss of 0.07 for one incorrect omission; for $R = 20$ there is no such loss. With $\delta = 0.4$, there are many false discoveries (15 and 22), but their count is not part of the criterion for the quality of the method.

The second block (Method B1) gives the losses with the Bayesian method based on the prior distribution comprising components $\mathcal{N}(0, 0.01)$, $\mathcal{N}(-1, 0.04)$, and $\mathcal{N}(1, 0.0625)$ with respective probabilities $p_O = 0.98$ and $p_L = p_U = 0.01$. The deviation of this prior from the model by which the data $\{\hat{\theta}_d\}$ was generated is deliberate, to mimic a realistic scenario. The method yields smaller loss than Method F for $\delta = 0.4$ (1.66 vs. 2.68 for $R = 10$ and 3.06 vs. 4.50 for $R = 20$), but greater loss for $\delta = 0.5$. Method B2 uses the prior distribution, by which the data was generated—the ordinary units as a mixture of the constant zero (78%) and $\mathcal{N}(0, 0.01), 20\%$, and the exceptional units as $\mathcal{N}(-1, 0.04)$ and $\mathcal{N}(1, 0.0625)$ with respective percentages 1.4 and 0.6. The classification is slightly poorer than for B1 with $\delta = 0.4$, and slightly better with $\delta = 0.5$.

Reducing the range of plausible penalty ratios would make little difference, although the one exceptional unit in the gray zone is close to the borderline. Reducing the range of plausible values of $\delta$, and raising its lower limit in particular, would reduce the number of equivocally classified ordinary units.
The classification with $\delta = 0.5$ is more successful than with $\delta = 0.4$. For the absolute and quadratic kernels with the same setting of $\delta$ and $R$, we obtained very similar results. Details are omitted.

If the standard errors were known, 0.08 for ordinary and 0.10 for exceptional units, but this information was not used to classify the units directly, the classifications would change only in a few cases, and the losses would be changed only slightly. There are no changes for Method B2. Thus, the uncertainty about the standard errors $s_d$ can be ignored. We have experimented with a wide range of settings of the priors and obtained the same result for most of them. The two incorrect omissions arose in all realistic settings. The classification seems to be more sensitive to the value of $\delta$ than to prior distributions for the ordinary and exceptional units.

The value of the prior information cannot be assessed reliably from one replication. We conducted a simulation study in which we replicated the study 500 times. That is, we replicated the processes of generating samples of sizes 1000 and 100 from the respective distributions of the non-zero ordinary and exceptional units, mimicked the process of their estimation, together with estimation of the associated standard errors, and applied the three methods of classification. The right-hand part of Table 1 displays the losses averaged over the replications. It shows that even imprecise prior information about $\{\theta_d\}$ is useful, and some further modest gains are attained by the exact prior. For example, the mean loss with Method F and $\delta = 0.4$ and $R = 10$ is 2.89. With the imprecise prior (Method B1) it is 2.04, and with the exact prior it is 1.77. The reductions are not proportional within the components $Q_X$ and $Q_O$; for each of the four settings $(\delta, R)$, the smallest values of $Q_O$ are attained by Method F. The discord between the single-replication and average losses in Table 1 suggests that the losses vary substantially across replications. The simulation study takes about 1 hour of CPU time using a custom-written function in R.

### 5. Analysis of a microarray experiment

We reanalyze the data of Alon and others (1999) on gene expression in tumor and normal colon tissues. The dataset comprises a matrix of 2000 genes (rows) and 62 tissue samples (columns), of which 40 are affected by tumor and 22 are normal; 22 patients supplied a tumor and a normal tissue sample each; and 18 further patients supplied one tumor tissue sample each. The dataset can be downloaded from [http://microarray.princeton.edu/oncology](http://microarray.princeton.edu/oncology).

The 2000 gene-specific $t$ statistics for comparing the two kinds of tissue samples are plotted in Figures 3(A) and (B). The statistics in Figure 3(A) compare the samples of sizes 40 and 22, and the statistics in Figure 3(B) are for the sets of 22 within-patient contrasts. The former statistics are associated with 60 d.f. each and the latter with 21 d.f. each. The densities of the $t$ distributions with these d.f. are drawn in each panel. Positive contrasts dominate in Figure 3(A); their majority in Figure 3(B) is much smaller. Figure 3(C) compares the sets of statistics more directly and confirms the systematic differences underlying the comparisons. We analyze the two sets of contrasts separately, to explore the impact of d.f., because we conjecture that the pairing within the patients is not relevant. Figure 3(D) contains the plot of the two sets of sample standard deviations (on the multiplicative scale). They are in a very wide range, dismissing the hypothesis that the underlying values might be constant and that their estimation could be pooled.

We apply the linear kernel with penalty ratios $R \in (1, 25)$. In view of the substantial dispersion of the standard deviations $\sigma$, we set $\delta$ to a fixed multiple of $\sigma$, $\delta = \lambda \sigma$, and regard $\lambda \in (2.5, 5.0)$ as plausible. The results are presented in Figure 4 in the form of contour plots of the numbers of genes classified as having differential expression and the total of the expected losses. The diagram shows that the specification of the plausible values of $\lambda$ and $R$ is essential because the number of selected genes is in a very wide range. The number decreases with $\lambda$ and increases with $R$; in fact, even the sets of selected genes decrease with $\lambda$ and increase with $R$. The numbers of selected genes are greater for the within-pair contrasts than for the contrasts of tissue types, except for the upper left-hand corner (large values of $\lambda$ and small values of $R$).
where the counts are smallest. The expected loss is uniformly greater for the within-patient contrasts, by between 12% and 72%. The percentage increases with $\lambda$ and decreases, more slowly, with $R$.

The number of selected genes is not a good estimator of the number of expressed genes because we cannot pretend that our selection is correct; the appropriate claim is that given the costs of the two kinds of error, the selection has minimum expected loss in total. We estimate the number of expressed genes by adding up the probabilities $p = P(|t| > \lambda)$ over the genes. These estimates are drawn in Figures 4(A) and (C) by thick dashes. For comparing the samples of tumor and normal tissues (40 vs. 22), the estimates are 349 for $\lambda = 2.5$ and 15.5 for $\lambda = 5.0$; the corresponding values for the within-patient contrasts are 403 and 16.3 for $\lambda = 2.5$ and 5.0, respectively. Assuming that the statistics $t$ are independent, the variance of this estimator is estimated by the total of $p(1 - p)$ over the 2000 genes. The corresponding (95%) confidence intervals are indicated in Figures 4(A) and (C) by thin dashes. The ratios of the estimates and
the associated standard errors are nearly 30 for the largest counts (at \( \lambda = 2.5 \)) and nearly 5.0 for the smallest counts (at \( \lambda = 5.0 \)).

If a target number of selected genes is specified, we assign a loss (cost) for exceeding this target as an additional (additive) component of the overall loss. As an example, suppose that every selected gene additional to the target of \( T \) is associated with cost \( \kappa > 0 \). If the selected set is greater than the target size, then we replace the original criterion for selection, the inequality \( Q_X < Q_O \), by the inequality \( Q_X + \kappa < Q_O \). If this leads to a selected set smaller than the target, then \( \kappa \) in this inequality is reduced to match the target size. A unit cost for too small a selection can be set and applied similarly. By applying a constraint on the size the expected loss is increased.

Figure 5 displays the results for the setting indicated in the caption. With increasing cost factor \( \kappa \) the number of selected genes decreases toward the target of 200, but the expected cost increases. For \( R = 5 \) and \( \lambda = 3 \), the target selection size is achieved for \( \kappa = 0.35 \) for the contrasts of tissue types (40 vs. 22,
6. Conclusion

Classification of units based on the values of a variable that are subject to estimation or measurement error, or are replaced by a proxy variable is one of the basic problems in statistics, with applications in gene expression studies, fraud detection, screening and performance assessment of institutions. We introduced a method that incorporates the consequences of the two kinds of incorrect classification by means of asymmetric loss functions. The asymmetry is characterized by the penalty ratio. Together with the kernel of the loss, they have to be informed by how the results of the analysis will be used, presuming that for each unit there are only two options: to treat it as if it were ordinary or exceptional.

We regard all alternatives in which a hypothesis is tested for each unit as unsatisfactory, because the consequences of the incorrect classification cannot be incorporated in a test, and a failure to reject a null hypothesis is illogically interpreted as its acceptance, leading to the unsupportable conclusion that the unit is ordinary. The assumption of normality of the within-class distributions may be restrictive, but substantial generalization is achieved by using mixtures of normal distributions. The evaluations for $t$- and $F$-distributed statistics are similar. Using the exponential family of distributions is an alternative avenue that still requires some exploration. More realistic specification of the loss function(s), informed by the subsequent agenda, and implementation of constraints that span two or more distinct markers $\theta_d$ present further challenges.

The computation for the Bayesian analysis is somewhat more complex, but even the example with 5000 units is analyzed within a few seconds of CPU time. The frequentist method involves iterations (NR algorithm) used for evaluating the borderline function $g(s)$ which separates the units classified as ordinary and exceptional. The algorithm converges very fast and the amount of computing is a fraction of the time for the Bayesian analysis. Computational speed is essential for a sensitivity study, in which the data is reanalyzed with alternative plausible assumptions.
In the analysis of gene expression data, we highlighted the difference between the count of the genes classified as expressed and estimation of the number of expressed genes. The selection errs on the side of false positives when such errors are less costly than false negatives. In contrast, this factor is not relevant for estimating the number of expressed genes. The wide range of counts in Figure 4 confirms that the penalty ratio (and the loss function in general) and the meaning of “small” (unimportant) in gene expression are crucial inputs to the analysis, to be elicited from the party with a stake in the analysis.

7. Software
Software in the form of R code is available on request from the author (sntlnick@sntl.co.uk).

Supplementary Material

Acknowledgments
Research for this article was partly supported by the grant ECO2011-28875 from the Spanish Ministry of Science and Innovation. Conflict of Interest: None declared.

References


[Received April 2, 2013; revised July 31, 2013; accepted for publication July 31, 2013]