Biometrika (2015), **102**, 4, *pp.* 753–766
Printed in Great Britain

doi: 10.1093/biomet/asv050
Advance Access publication 4 November 2015

Optimal multiple testing under a Gaussian prior on the effect sizes

By EDGAR DOBRIBAN

Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.
dobriban@stanford.edu

KRISTEN FORTNEY, STUART K. KIM

Department of Developmental Biology, Stanford University, Stanford, California 94305, U.S.A.
kfortney@stanford.edu stuartkm@stanford.edu

AND ART B. OWEN

Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.
owen@stanford.edu

Summary

We develop a new method for large-scale frequentist multiple testing with Bayesian prior information. We find optimal p-value weights that maximize the average power of the weighted Bonferroni method. Due to the nonconvexity of the optimization problem, previous methods that account for uncertain prior information are suitable for only a small number of tests. For a Gaussian prior on the effect sizes, we give an efficient algorithm that is guaranteed to find the optimal weights nearly exactly. Our method can discover new loci in genome-wide association studies and compares favourably to competitors. An open-source implementation is available.

Some key words: Genome-wide association study; Multiple testing; Nonconvex optimization; p-value weighting; Weighted Bonferroni method.

1. Introduction

The research presented in this paper is motivated by the genetics of human longevity. Genome-wide association studies of longevity compare long-lived individuals with matched controls (Brooks-Wilson, 2013). More than 500 000 genetic variants have been tested for their association with longevity, which amounts to a large multiple hypothesis testing problem. In addition to multiplicity, the sample size is small, usually of the order of a few hundred. As a consequence, only a few loci have been replicably associated with human longevity, and they do not explain the heritability of the trait (Hjelmborg et al., 2006).

The multiplicity may be countered by testing only a few candidate variants selected based on prior scientific knowledge. In a separate work in preparation, led by the second author, we find that a more general genome-wide test helps to improve power in a study of longevity. We leverage prior information from genome-wide association studies of age-related diseases, such as coronary artery disease and diabetes. For this task, we develop a new large-scale method of frequentist multiple testing with Bayesian prior information. In this paper we provide the theory for this method.

Our method is a novel p-value weighting scheme; p-value weighting is a general methodology for multiple testing that leverages independent prior information to improve power (Roeder & Wasserman, 2009; Gui et al., 2012). Suppose that we test hypotheses H_i via the p-values P_i for $i = 1, \ldots, J$. For a significance level $q \in [0, 1]$, the weighted Bonferroni method declares the ith hypothesis to be significant if $P_i \leqslant q w_i$. The weights $w_i \geqslant 0$ are based on independent data. The familywise error rate, the probability of making at least one error, is controlled if the weights average to 1, as it equals at most $\alpha = Jq$.

In previous work, optimal weights have been found in a Gaussian model of hypothesis testing. Let the test statistics in the current study be $T_i \sim N(\mu_i, 1)$, where the μ_i are the means, or effect sizes; we test the null hypotheses $\mu_i \geqslant 0$ against $\mu_i < 0$. We have some information about the μ_i from prior studies. Roeder & Wasserman (2009) and Rubin et al. (2006) considered a model where the μ_i are known exactly from the prior data, and the weights are allowed to depend on the μ_i. In such a model they found the optimal weights for the weighted Bonferroni method, which maximize the expected number of discoveries. We show that this amounts to solving a convex optimization problem.

The assumption that the μ_i are known precisely is problematic: if they were known, there should be no follow-up study. In practice, empirical estimates of the μ_i are used. However, the fixed-μ_i weights do not take into account the uncertainty in the estimates. Instead, we account for uncertainty explicitly by considering the model with uncertain prior information in the form $\mu_i \sim N(\eta_i, \sigma_i^2)$. Only the prior means η_i and standard errors σ_i are known from independent data, not the precise effect sizes. Finding the optimal weights, which we call Bayes weights, is then a nonconvex optimization problem.

Westfall et al. (1998) formulated a general framework that includes this problem as a special case and allows, for instance, for Student t-distributed priors. They used a direct numerical solver, a quasi-Newton optimization method which generally scales as $O(J^3)$, to find the weights. Published examples using this approach are typically small (Westfall et al., 1998; Westfall & Soper, 2001). This method of computing the weights does not scale up for our problems, which involve more than 500 000 genetic variants. Further, the generic quasi-Newton method has no guarantee of finding the global optimum of the nonconvex problem.

Our key contribution here is to provide an efficient method of finding the weights that maximize average power for the weighted Bonferroni method, in the model with Gaussian priors. We solve the optimization problem exactly for small q, less than a problem-dependent value which is often between 0·1 and 0·3. For larger q, we can solve the problem for a nearby q^* such that $|q^* - q| \leqslant 1/(2J)$. The cost per iteration of our algorithm is $O(J)$ in the first case and $O(J \log J)$ in the second case. We observe that a nearly constant number of iterations is used, regardless of J. We find it remarkable that this problem admits a near-exact solution.

For large-scale problems, this approach leads to a method for multiple testing that controls a frequentist error measure while also taking into account Bayesian prior information. This method follows George Box's advice to be Bayesian when predicting but frequentist when testing (Box, 1980). Similar ideas were used previously by Carlin & Louis (1985); see § 2. As mentioned, a more general formulation was also considered in Westfall et al. (1998) and Westfall & Soper (2001). We show that our approach is feasible for large-scale problems.

When prior information is uncertain, we show via simulations that the new method has more power and is more stable than competitors. We also show theoretically that weighting leads to substantially improved power. We apply the method to genome-wide association studies.

By analysing several such datasets, we show that our method has advantages in terms of power and easier tuning compared to other methods.

With rapidly increasing volumes of data available as prior information for any given study, our method should be useful for other problems in biology and elsewhere. The data analysis and computational results in this paper are reproducible, and an open-source implementation of the method is available from the authors.

2. Related work

There is a large literature on statistical methods for multiple testing with prior information, some of which is reviewed in Roeder & Wasserman (2009) and Gui et al. (2012). Spjøtvoll (1972) devised optimal single-step multiple testing procedures maximizing average or minimal power and controlling the familywise error rate. Later it was recognized that Spjøtvoll's results are equivalent to optimal p-value weighting methods. For instance, Benjamini & Hochberg (1997) developed extensions of Spjøtvoll's methods for p-value weighting, allowing for weights also in the importance of the hypotheses.

Leveraging Spjøtvoll's results, Rubin et al. (2006) and Roeder & Wasserman (2009) found an explicit formula for optimal weights of the weighted Bonferroni method in the Gaussian model $N(\mu_i, 1)$, assuming that the effects are known exactly. In practice the effects are estimated, but the weights do not take this into account. These weights are optimal for average power, and this efficient method is suitable for large applications. Eskin (2008) and Darnell et al. (2012) applied the framework of Roeder & Wasserman (2009) to genome-wide association studies; they accounted for correlations between the tests but assumed that the effects are known exactly.

Another popular approach is to test the top candidates from a prior study, often known as two-stage testing or candidate study. It can be viewed as a p-value weighting method where some of the weights equal zero. A specific version for genome-wide association studies has been called the proxy-phenotype method (Rietveld et al., 2014).

In the literature on carcinogenicity trials, related methods have been devised to select tumour sites based on historical data (Carlin & Louis, 1985; Louis & Bailey, 1990); the methods are explicitly Bayesian with regard to historical data and frequentist in analysing current data. These models and methods differ from ours, and focus on pairwise comparisons based on Fisher's exact test (Louis & Bailey, 1990).

Westfall et al. (1998) considered a Gaussian model $N(\mu_i, 1)$ for the effects in hypothesis testing where prior distributions are known for the means. They formulated the problem of finding the weights that maximize expected power for the weighted Bonferroni method, and this was followed up for binary data in Westfall & Soper (2001), motivated by carcinogenicity trials. As mentioned in § 1, published studies using their optimization methods are typically small.

Less work exists on weighted methods beyond the single-step Bonferroni method, or beyond the control of the familywise error rate. The step-down method of Holm (1979) can use weights, and Westfall & Krishen (2001) and Westfall et al. (2004) discuss the choice of optimal weights. Genovese et al. (2006) showed that the weighted Benjamini–Hochberg procedure controls the false discovery rate, and Roquain & Van De Wiel (2009) proposed a method of choosing weights optimally, assuming fixed known effects. Peña et al. (2011) developed a general framework for optimal multiple decision functions for the control of familywise error rate and false discovery rate, assuming exact knowledge of the alternatives.

In this paper we focus on the familywise error rate, because it is the standard measure of error controlled in our motivating application, and because in this case it is already challenging to find the optimal weights accounting for uncertainty on a large scale. Extension of this work to the Benjamini–Hochberg procedure and to false discovery rate control is left for future research.

3. THEORETICAL RESULTS

3·1. *Background*

We work in the Gaussian means model of hypothesis testing: we observe test statistics $T_i \sim N(\mu_i, 1)$ and test each null hypothesis $H_i : \mu_i \geqslant 0$ against $\mu_i < 0$. The p-value for testing H_i is $P_i = \Phi(T_i)$, where Φ denotes the normal cumulative distribution function.

For a weight vector $w \in [0, \infty)^J$ and a significance level $q \in [0, 1]$, the weighted Bonferroni procedure rejects H_i if $P_i \leqslant q w_i$. Usually this corresponds to $w_i = 1$. For general weights, the expected number of false rejections, known as the per-family error rate, equals $\sum_{\mu_i \geqslant 0} \mathrm{pr}(P_i \leqslant q w_i) = q \sum_{\mu_i \geqslant 0} w_i$. If $\sum_{i=1}^J w_i \leqslant J$, the expected number of false rejections is at most $\alpha = Jq$. By Markov's inequality, this implies that the familywise error rate is at most α. Hence the weighted Bonferroni method controls the familywise error rate. This result does not require independence of the T_i. We assume always that $q \leqslant 1$, and usually that $q \ll 1$. Without loss of generality, we restrict the weights to the interval $[0, 1/q]$.

Let us denote the number of rejections by $R(w) = \sum_{i=1}^J I(P_i \leqslant q w_i)$, where $I(\cdot)$ is the indicator function. The optimal weights maximizing the expected number of discoveries, assuming a priori known effects μ_i, were found explicitly by Roeder & Wasserman (2009) and independently by Rubin et al. (2006). Denoting by $E_T(\cdot)$ the expectation with respect to T_i, they solved the constrained optimization problem

$$\max_{w \in [0, 1/q]^J} E_T\{R(w)\} \quad \text{subject to} \quad \sum_{i=1}^J w_i = J. \tag{1}$$

It was not noted previously that this problem is convex. The objective is a sum of terms of the form $\Phi\{\Phi^{-1}(q w_i) - \mu_i\}$, whose concavity follows directly by differentiation. Yet, by simple Lagrangian optimization, the above papers showed that if all $\mu_i < 0$, the optimal weights are $w_i = w(\mu_i)$ where

$$w(\mu) = \Phi\left(\frac{\mu}{2} + \frac{c}{\mu}\right) \Big/ q. \tag{2}$$

Here c is the unique normalizing constant such that the weights sum to J. Interestingly, the weights are not monotonic as a function of μ, but are largest for intermediate values of μ. As noted by Roeder & Wasserman (2009), formula (2) is a direct consequence of Spjøtvoll's theory of optimality in multiple testing (Spjøtvoll, 1972). Accordingly, we call these weights the Spjøtvoll weights.

3·2. *Weighting leads to substantial power gain*

To illustrate theoretically that p-value weighting can lead to increased power, we compare the power of optimal weighting with that of unweighted testing in a sparse mixture model.

First, we note that p-value weighting exploits the heterogeneity of the tests. In the simplest case there are only large and small negative effects, say $M \ll m \approx 0$. We consider the $m \approx 0$ limit, and for simplicity we suppose that $m = 0$. Let the fractions of large and small effects be $\pi_1 > 0$ and $\pi_0 > 0$, respectively, so that $\pi_1 J$ of the means equal M and the remaining $\pi_0 J$ equal zero. We solve for the optimal weights.

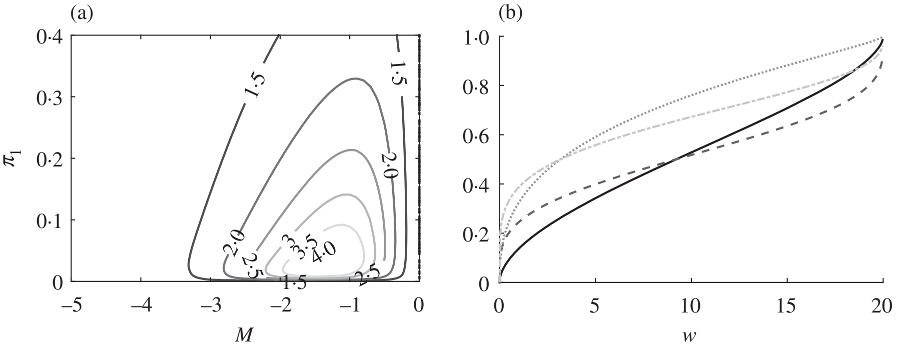

Fig. 1. Power gain and nonconvexity: (a) contour plot of the power ratio of optimal to unweighted testing for sparse means; (b) plots of four different instances of the function that is summed in the optimization objective; the nonconvex summand $w \mapsto \Phi[\{\Phi^{-1}(qw) - \eta\}/\gamma]$ with $q = 0{\cdot}05$ is plotted for the pairs $(\eta, \sigma) = (-0{\cdot}1, 1)$ (solid), $(-0{\cdot}1, 2)$ (dashed), $(-1, 1)$ (dotted) and $(-1, 2)$ (dot-dashed).

Proposition 1. *There is a set of optimal p-value weights that gives the same weights to the same means, i.e., weights w_0 and w_1 to means* 0 *and M, respectively, where*

$$(w_0, w_1) = \begin{cases} (0,\ 1/\pi_1), & \pi_1\Phi(-|M|/2) > q, \\ \left(\dfrac{q - \pi_1\Phi(-|M|/2)}{q\pi_0}, \dfrac{\Phi(-|M|/2)}{q}\right), & \pi_1\Phi(-|M|/2) \leqslant q. \end{cases}$$

Further, the power of the optimal p-value weighting method is

$$p^*(\pi_1, M, q) = \begin{cases} \pi_1\Phi\{\Phi^{-1}(q/\pi_1) + |M|\}, & \pi_1\Phi(-|M|/2) > q, \\ q + \pi_1\{\Phi(|M|/2) - \Phi(-|M|/2)\}, & \pi_1\Phi(-|M|/2) \leqslant q. \end{cases}$$

If the absolute effect size $|M|$ is small enough that $\pi_1\Phi(-|M|/2) > q$, all the weight is placed on the larger means, which is the behaviour we would expect intuitively. However, if $|M|$ is large enough that $\pi_1\Phi(-|M|/2) \leqslant q$, then it is advantageous to place some weight on the small means, because a large absolute effect size $|M|$ will be detected with high probability.

In Fig. 1(a) we plot the ratio $p^*(\pi_1, M, q)/p(\pi_1, M, q)$ for $q = 10^{-3}$, where $p(\pi_1, M, q) = \pi_0 q + \pi_1\Phi\{\Phi^{-1}(q) + |M|\}$ is the power of unweighted Bonferroni testing. For most effect sizes $M \in [-2{\cdot}5, -0{\cdot}25]$ and for $\pi_1 < 0{\cdot}4$, we see a power gain of at least 50% relative to unweighted Bonferroni testing. Moreover, there is a hotspot where the power gain can be three- to four-fold. Optimal weighting can lead to a significant gain in power.

3·3. *Weights with imperfect prior knowledge*

In the previous sections it was assumed that the effects μ_i are known precisely. We now assume that we have uncertain prior information in the form $\mu_i \sim N(\eta_i, \sigma_i^2)$.

Following Westfall et al. (1998), we maximize the expected power $E_\mu[E_T\{R(w)\}]$ averaged with respect to the random T_i and μ_i. Introducing $\gamma_i = (\sigma_i^2 + 1)^{1/2}$, the optimization problem, which we call the Bayes weights problem, becomes

$$\max_{w \in [0, 1/q]^J} \sum_{i=1}^{J} \Phi\left\{\frac{\Phi^{-1}(qw_i) - \eta_i}{\gamma_i}\right\} \quad \text{subject to} \quad \sum_{i=1}^{J} w_i = J. \tag{3}$$

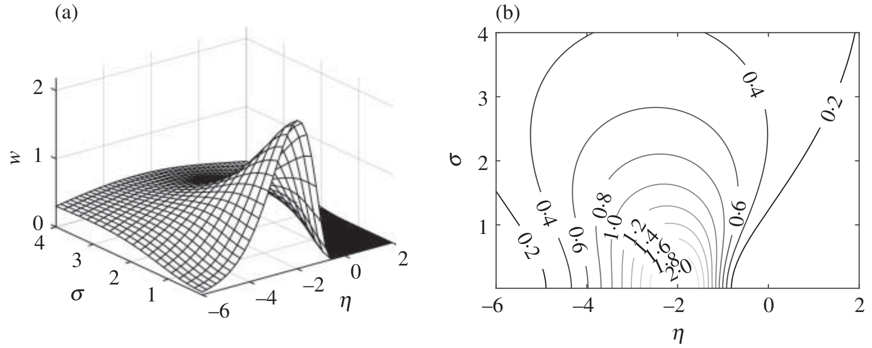

Fig. 2. Bayes weights: (a) surface plot and (b) contour plot of the Bayes weight function $w(\eta, \sigma)$ defined in Theorem 1; Spjøtvoll weights are on the segment $\{\sigma = 0, \eta < 0\}$.

This objective function is not concave if any $\gamma_i > 1$. To help with visualization, the function $w \mapsto \Phi[\{\Phi^{-1}(qw) - \eta\}/\gamma]$ is plotted in Fig. 1(b) for four parameter pairs (η, γ). On the interval $w \in [0, 1/q]$, the function is first concave and then convex.

Our main contribution is to solve this problem efficiently for large J. The results in this respect are two-fold. First, we can solve the problem exactly in the special case where q is sufficiently small. Second, we have a nearly exact solution for arbitrary q. Starting with the simpler first case, we define

$$c(\eta, \gamma; \lambda) = -\frac{\eta + \gamma\{\eta^2 + 2(\gamma^2 - 1)\log(\gamma\lambda)\}^{1/2}}{\gamma^2 - 1}.$$

A weighted one-sided test $\Phi(T_i) \leqslant qw_i$ can be written equivalently in terms of the critical values as $T_i \leqslant \Phi^{-1}(qw_i)$. It turns out that the critical values corresponding to the optimal Bayes weights can be expressed in terms of c, when q is small enough that

$$q \leqslant \frac{1}{J}\sum_{i=1}^{J} \Phi\{c(\eta_i, \gamma_i; 1)\}. \tag{4}$$

In our data analysis examples and simulations, this mild restriction requires that q be below values in the range 0·1 to 0·3. In the next result we give the exact optimal weights for small q when all $\sigma_i > 0$.

Theorem 1. *If the significance level $q \in [0, 1]$ is small enough that (4) holds, then the optimal Bayes weights maximizing the average power (3) are $w_i = w(\eta_i, \gamma_i; \lambda) = \Phi\{c(\eta_i, \gamma_i; \lambda)\}/q$, where $\lambda \geqslant 1$ is the unique constant such that $\sum_{i=1}^{J} w(\eta_i, \gamma_i; \lambda) = J$.*

In the Supplementary Material, we solve this problem by maximizing the Lagrangian. Two key properties that we use are joint separability of the objective function and constraint, and analytic tractability of the Gaussian density.

Figure 2 displays an instance of the optimal weights $w(\eta, \sigma)$ as a function of the prior mean η and the standard deviation σ. In the theorem the weights are a function of (η, γ), but they can also be viewed as a function of (η, σ) via the natural map $\gamma^2 = \sigma^2 + 1$. As the standard error σ becomes small, our weights tend to the Spjøtvoll weights.

Proposition 2. *For any λ and $\eta < 0$, the Bayes weight function defined by $w(\eta, \gamma; \lambda) = \Phi\{c(\eta, \gamma; \lambda)\}/q$ tends to the Spjøtvoll weight function defined in (2) as $\sigma \to 0$.*

With $\sigma_i > 0$, the weights are regularized: more extreme weights are shrunk towards a common value in a nonlinear way. For finite σ_i, our weights can be viewed as a smooth interpolation between Spjøtvoll weights and uniform weights. It is reasonable to think at first that as all $\sigma_i \to \infty$, the best weight allocation becomes the uniform one. However, this is not the case: a symmetry-breaking phenomenon occurs due to nonconvexity.

Consider a weight vector w that equals $1/q$ for $\lfloor Jq \rfloor$ indices, and assume that Jq is not an integer. Distribute the remaining strictly positive weight equally among the remaining hypotheses. It is now easy to see that the hypotheses with weights equal to $1/q$ are always rejected, so their power equals 1. For the remaining hypotheses the power $\Phi[\{\Phi^{-1}(qw) - \eta\}/\gamma]$ tends to $\Phi(0) = 1/2$ as $\gamma \to \infty$. This shows that the limiting power of this unbalanced weighting scheme is $\lfloor Jq \rfloor + (J - \lfloor Jq \rfloor)/2 = (J + \lfloor Jq \rfloor)/2$. For uniform weighting, the power tends to $1/2$ as $\gamma \to \infty$, for each hypothesis. This shows that the limiting power of uniform weighting is $J/2$. Hence, the power of the skewed weighting scheme is larger than that of uniform weighting. This illustrates the symmetry-breaking phenomenon caused by the extreme nonconvexity of the optimization problem.

Fortunately, the situation is better when condition (4) holds. In addition to being easy to check for any given parameters q and (η_i, σ_i), we now show that the constraint is mild. Often we want to keep $\alpha = Jq$ small even if J is large, because α is the number of false rejections that we tolerate. In this regime, the condition holds as long as there are a few average-sized negative prior means η_i. We denote by $z_c = \Phi^{-1}(c)$ the normal quantile function.

PROPOSITION 3. *Condition* (4) *holds if there are $K \geqslant 1$ distinct indices i with negative η_i, for which $\gamma_i^2 \log(\gamma_i^2)/|z_{\alpha/K}| \leqslant |\eta_i| \leqslant |z_{\alpha/K}|$.*

If $\alpha/K \to 0$, then $|z_{\alpha/K}| \simeq \{2\log(K/\alpha)\}^{1/2}$, so the simple condition holds provided that $\gamma_i^2 \log(\gamma_i^2)\{2\log(K/\alpha)\}^{-1/2} \lesssim |\eta_i| \lesssim \{2\log(K/\alpha)\}^{1/2}$. For instance, if $K = 10$ and $\alpha = 0{\cdot}01$, then $\{2\log(K/\alpha)\}^{1/2} \simeq 3{\cdot}7$. If, moreover, $\sigma = 1$ so that $\gamma^2 = 2$, and $\gamma^2 \log(\gamma^2)/3{\cdot}7 = 0{\cdot}16$, then we need only ten effect sizes with $0{\cdot}16 \leqslant |\eta| \leqslant 3{\cdot}7$. This is a weak requirement.

When q is small, we use a damped Newton's method to find the right constant λ from Theorem 1 via a one-dimensional line search. The function evaluations cost $O(J)$ per iteration, and empirically we find that the algorithm takes only a small number of iterations to converge, independently of J. We can solve problems involving more than two million tests in a few seconds on a desktop computer.

Now we present our result for the general case.

THEOREM 2. *For any $q \in [0, 1]$, the nonconvex Bayes weights problem can be solved for a nearby $q^* \in [0, 1]$ for which $|q^* - q| \leqslant 1/(2J)$. The optimal weights and q^* can be found in $O(J \log J)$ steps.*

This result is relevant when α, the expected number of errors under the null hypothesis, is controlled at a threshold greater than 1/2. Our weights will be optimal for a q^* that is close to q. We see from the proof that even for large q, q^* often equals q. The method also returns the value of q^*, which the user can inspect. It is then the user's decision as to whether to perform multiple testing adjustment at the original level q or at the new level q^*.

The analysis of nonconvex optimization problems is challenging. It seems remarkable that the nonconvex Bayes weights problem admits a nearly exact solution.

4. Simulation studies

4·1. *Bayes weights are more powerful than competing weighting schemes*

We perform two simulation studies to explore the empirical performance of our method. First, we show that Bayes weights increase power more reliably than two other weighting schemes, namely exponential weights and filtering.

For Bayes weights, we multiply the variances by a dispersion factor ϕ, i.e., $N(\eta_i, \phi\sigma_i^2)$. The default value for this tuning parameter is $\phi = 1$ and, as discussed in § 5·3, we recommend use of the default value in most cases. The purpose of changing the dispersion is to explore the robustness of our method with respect to misspecification of the prior variances. The dispersion ranges from 0 to 4, and Spjøtvoll weights correspond to $\phi = 0$.

Exponential weights with tilt parameter β are defined as $w_i = \exp(\beta|\eta_i|)/c$, where $c = \sum_{i=1}^{J} \exp(\beta|\eta_i|)/J$. This weighting scheme was proposed by Roeder et al. (2006), who recommend $\beta = 2$ as the default value. We consider the range $\beta \in [0, 4]$. As noted by Roeder et al. (2006), exponential weights are sensitive to large means. To guard against this sensitivity, we truncate weights larger than $1/q$ and redistribute their excess weight among the next largest weights.

Filtering methods test only the most significant effects $\eta_i \leqslant M$, using the unweighted Bonferroni method. These methods can be viewed as weighting schemes in which some weights are zero. Such methods are known under many names, such as two-stage testing, screening, or proxy-phenotype methods (Rietveld et al., 2014). We adopt the term filtering used by Bourgon et al. (2010), who filter based on independent information in the current dataset rather than prior information. The threshold M ranges from -4 to 0. If $|M|$ is large and fewer than Jq hypotheses would be tested, then we instead test the most significant Jq hypotheses.

In the simulation, we generate $J = 1000$ random means and variances independently according to $\eta_i \sim N(0, 1)$ and $\sigma_i \sim |N(0, 1)|$, and we set $q = 10^{-2}$. For any weight vector w, we calculate the power as the objective from (3) divided by J, to reflect the average power per test.

The results are shown in Fig. 3(a). Each method can improve the power over unweighted testing. However, Bayes weights yield more power than the other methods. The best power is attained when the dispersion ϕ is equal to 1, but good power is reached in a large neighbourhood of $\phi = 1$. Our weights are robust with respect to misspecification of the tuning parameter.

In particular, taking uncertainty into account helps. Spjøtvoll weights, which assume fixed and known effects, and are represented on the figure as regularized weights with $\phi = 0$, have less power than Bayes weights with positive ϕ, for a wide range of ϕ.

The remaining two methods, filtering and exponential weights, have disadvantages. While filtering yields a gain in power for a thresholding parameter $M \lesssim 3/4$, it also leads to a substantial power loss for $M > 1$. For sufficiently large M the power equals q, because only the top Jq hypotheses are selected. Another significant disadvantage is that there seems to be no principled way to choose M a priori without additional assumptions. Similarly, exponential weighting leads to at most a small gain in power, and it usually leads to a power loss.

We conclude that Bayes weights are robust with respect to the choice of the tuning parameter and have uniformly good power. In contrast, exponential weighting and filtering are more sensitive, and their power can drop substantially.

4·2. *Bayes weights have a worst-case advantage*

We show that Bayes weights have a worst-case advantage compared to Spjøtvoll weights. We use the sparse means model and generate $J = 1000$ means η_i distributed as $\eta_i \sim \pi_0\delta_m + \pi_1\delta_M$,

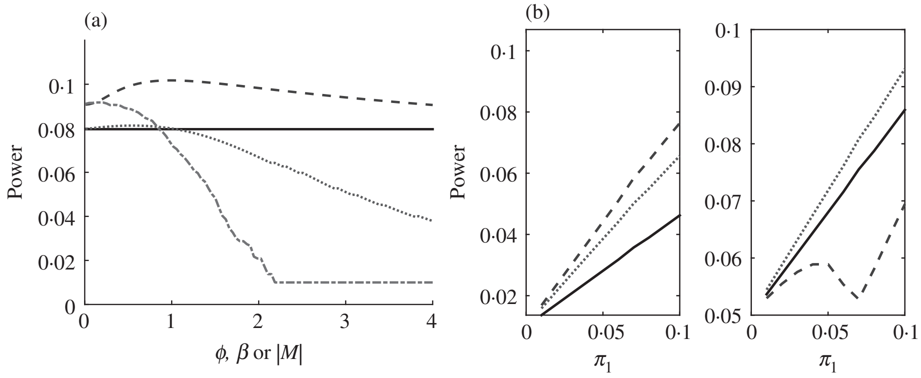

Fig. 3. (a) Power of four p-value weighting methods plotted as a function of their parameter: unweighted (solid), Bayes (dashed) as a function of the dispersion ϕ, exponential (dotted) as a function of β, and filtering (dot-dashed) as a function of $|M|$; the Spjøtvoll weights correspond to the point at the origin $\phi = 0$ on the Bayes weights curve. (b) Power comparison for sparse means: deterministic (left) and average (right) power plotted as a function of the proportion of large means π_1, for the unweighted (solid), Spjøtvoll (dashed) and Bayes (dotted) methods.

where $m = -10^{-3}$ and $M = -2$. We set $q = 10^{-2}$ and vary π_1 from 0 to 0·1. We set all σ_i to equal σ and consider $\sigma = 0$ or 1.

Spjøtvoll weights are optimal for $\sigma = 0$, while Bayes weights are optimal for $\sigma = 1$. We evaluate these weighting schemes by calculating the power that they do not maximize, i.e., the average power (3) for Spjøtvoll weights and the deterministic power (1) for Bayes weighting. We also compute the power of the unweighted Bonferroni method.

The results are displayed in Fig. 3(b). Bayes weights lose only a little power compared to the optimal Spjøtvoll weights. In contrast, Spjøtvoll weights lose a lot of power relative to Bayes weights, which maximize the worst-case power. Bayes weights show a maximin property. Further, as shown in the Supplementary Material, Spjøtvoll weights lose power near $\pi_1 = 0{\cdot}07$ because they set the weights equal to zero on the small means.

5. Application to genome-wide association studies

5·1. *Review of genome-wide association studies*

We adapt our framework to genome-wide association studies, relying on basic notions of quantitative genetics (see, e.g., Lynch & Walsh, 1998). In this section we present in detail the methodology for this application, while also illustrating the steps of using our framework for specific problems.

We study a quantitative trait y in a population, with the goal of understanding the effects of single nucleotide polymorphisms $g_1, \ldots, g_J$ on the trait. We assume that y has mean 0 and known variance; here g_i denotes the centred minor allele count of variant i for an individual. We rely on the linear model for the effect of the ith variant on the trait: $y = g_i\beta_i + \varepsilon_i$. In this model y is the phenotype of a randomly sampled individual from the population, so g_i is random, β_i is a fixed unknown constant, and ε_i is the residual error. This error is a zero-mean random variable that is independent of g_i, with variance σ_i^2.

Suppose that we observe a sample of N independent and identically distributed observations from this model. We use the standard linear regression estimate $\hat{\beta}_i$, which for a large sample size has an approximate distribution $N^{1/2}\hat{\beta}_i \mathrel{\dot\sim} N\{N^{1/2}\beta_i, \sigma_i^2/\text{var}(g_i)\}$. To standardize, we divide by τ_i, where $\tau_i^2 = \sigma_i^2/\text{var}(g_i)$ is the variance of $N^{1/2}\hat{\beta}_i$.

With these steps, we have framed our problem in the Gaussian means model. Writing $T_i = N^{1/2}\hat{\beta}_i/\tau_i$ and $\mu_i = N^{1/2}\beta_i/\tau_i$, we have $T_i \mathrel{\dot{\sim}} N(\mu_i, 1)$, which has the required form. Let us also define the standardized effect size $\nu_i = \beta_i/\tau_i$, which will be of key importance.

5·2. *Prior information*

To use prior information, assume that we also have a prior trait y_0 which is measured independently on a different, independent sample from the same population. With the same assumptions on y_0, we can write $y_0 = g_i\beta_{0i} + \varepsilon_{0i}$. Here β_{0i} is a fixed unknown constant, and ε_{0i} is random. Suppose that we have independent samples of size N_i and N_{0i} for the two traits. If we define T_{0i} and ν_{0i} by analogy to the definitions for y, we can write $T_{0i} \sim N(N_{0i}^{1/2}\nu_{0i}, 1)$.

We model the relatedness of the two traits as a relation between the standardized effect sizes ν, which do not depend on the sample size. If the two traits are closely related, the first-order approximation is equality, or $\nu_i = \nu_{0i}$. This model captures the pleiotropy between the two traits (Solovieff et al., 2013).

The final step is to compute the distribution of ν_i given the prior data T_{0i}. For this we need to choose a prior for ν_i, and for simplicity we will use a flat prior.

We now have all ingredients for the model of Gaussian hypothesis testing with uncertain information. Specifically, we have $\mu_i \sim N(\eta_i, \sigma_i^2)$, where $\mu_i = N_i^{1/2}\nu_i$, $\eta_i = (N_i/N_{0i})^{1/2}T_{0i}$ and $\sigma_i^2 = N_i/N_{0i}$.

The uncertainty in T_{0i} may be different from 1, and may exceed it due to overdispersion. This is one way to weaken the first-order approximation assumption. To allow for overdispersion, we recall the parameter ϕ used in our simulation. We model the prior data as $T_{0i} \sim N(N_{0i}^{1/2}\nu_{0i}, \phi)$, and then the variance becomes $\sigma_i^2 = \phi N_i/N_{0i}$. The default value $\phi = 1$ is recommended in most cases. Finally, we compute the Bayes weights w_i with parameters q and (η_i, σ_i^2), and we run the weighted Bonferroni method on the current p-values. This fully specifies the method, which is summarized in the following algorithm.

Algorithm 1. Bayes-weighted Bonferroni multiple testing in genome-wide association studies

(i) Let T_{0i} be the prior effect sizes for $i = 1, \ldots, J$.
(ii) Let N_{0i} and N_i be the prior and current sample sizes.
(iii) Let P_i be the current p-values.
(iv) Let q be the significance threshold; the default value is $q = 0{\cdot}05/J$.
(v) Let ϕ be the dispersion; the default value is $\phi = 1$.
(vi) Set the prior means and variances: $\eta_i = (N_i/N_{0i})^{1/2}T_{0i}$ and $\sigma_i^2 = \phi N_i/N_{0i}$.
(vii) Compute the Bayes weights w_i, defined via (3), with parameters q and (η_i, σ_i^2).
(viii) Output indices i such that $P_i \leqslant q w_i$.

5·3. *Practical remarks*

It is important that we retain Type I error control even when the modelling assumptions fail. The only requirement is that we have marginally valid p-values. We list two common deviations from our model. First, summary data for genome-wide association studies sometimes include only the magnitude of the effects and not their sign. In this case we have two choices: we could assume that the directions of effects are the same, and perform a one-tailed test of the current effect in the prior direction; alternatively, we could do a two-tailed test by including the tests with prior parameters (η_i, σ_i^2) and $(-\eta_i, \sigma_i^2)$ for each i, for a total of $2J$ tests. Large effects will often be in the same direction, whereas small effects may change direction between the prior and current studies. Our procedure for dealing with two-sided effects may lead to minor power loss while retaining Type I error control. On the other hand, in some cases the prior and current traits can be of different types; for instance, the prior trait could be binary and the current trait quantitative. In such a situation, the model $\nu_i = \nu_{0i}$ should be re-examined, but it is still convenient to use as a first approximation.

We recommend using the default value of the tuning parameter, $\phi = 1$, in all but exceptional cases. This value was derived from a natural Bayesian model, and our simulations and data analysis show that it provides good performance in most cases. The same numerical results demonstrate that our method is not too sensitive to the choice of tuning parameter. If the relationship between the two traits is thought to be weak, one could use a larger ϕ, such as $\phi = 2$. If the uncertainty in the prior information is less than that suggested by the usual model, one could use a small ϕ, such as $\phi = 0{\cdot}5$. If the value $\phi = 1$ was tried first, the results of that analysis should also be reported.

One may wish to use the weighted Benjamini–Hochberg method with our weights (Genovese et al., 2006); but in general this will be underpowered, as optimal weights for stepwise methods differ greatly from those for single-step methods (Westfall & Soper, 2001). However, in the special case of very small q, in our data analysis examples we have observed that weights often become monotonically increasing with the magnitude of the effect size, and thus are similar to the optimal weights for stepwise methods.

6. Data analysis

6·1. *Data sources*

We illustrate the application of our method by analysing data from publicly available genome-wide association studies. We use the p-values, recorded for 500 000 to 2·5 million genetic variants, from five studies: CARDIoGRAM and C4D for coronary artery disease (Schunkert et al., 2011; Coronary Artery Disease Genetics Consortium, 2011), blood lipids (Teslovich et al., 2010), schizophrenia (Schizophrenia Psychiatric Genome-Wide Association Study Consortium, 2011), and estimated glomerular filtration rate creatinine (Köttgen et al., 2010); see the Supplementary Material.

We analyse three pairs of datasets, with a specific motivation for each. First, we use CARDIoGRAM as prior information for C4D. This is a positive control for our method, since both studies measure coronary artery disease. We choose C4D as the target because it has a smaller sample; hence prior information may increase power more substantially.

Second, we use the blood lipids study as prior information for the schizophrenia study. Andreassen et al. (2013) demonstrated improved power with this pair. They used a fully Bayesian method, and our goal is to evaluate the power improvement using a frequentist method. There is a small overlap between the controls of the two studies.

Third, we use the creatinine study as prior information for the C4D study. Heart disease and renal disease are comorbid (Silverberg et al., 2004), so this set-up may improve power.

6·2. *Methods and additional details*

We run weighted Bonferroni multiple testing for each of five weighting schemes. The prior data is $T_{0i} = \Phi^{-1}(P_{0i}/2)$, where P_{0i} is the ith prior p-value. The familywise error rate is controlled at $\alpha = 0{\cdot}05$, so that the p-value thresholds are approximately 10^{-9} to 10^{-8}.

Table 1. *Number of significant loci for five methods on three examples: the top portion of the table shows results pruned for linkage disequilibrium, the middle part shows results without pruning, and the bottom portion reports the score of each method*

			Bayes(ϕ)			Exp(β)			Filter($-\log P$)		
Parameter	Un	Spjot	0·1	1	10	1	2	4	2	4	6
Pruned											
CG → C4D	4	11	10	8	4	4	5	4	10	10	6
Lipids → SCZ	4	1	1	1	5	1	0	0	2	2	2
eGFRcrea → C4D	4	2	2	4	4	4	5	4	1	0	1
Unpruned											
CG → C4D	29	45	44	39	29	32	34	27	40	48	34
Lipids → SCZ	116	214	214	223	123	92	0	0	217	96	39
eGFRcrea → C4D	29	18	18	23	29	29	28	19	1	0	1
Scoring											
Score	0	0	0	1	1	0	0	−1	0	−1	−1
Total	0	0		sum = 2			sum = −1			sum = −2	

Un, unweighted; Spjot, Spjøtvoll; Bayes(ϕ), Bayesian with $\phi = 0{\cdot}1$, 1 or 10; Exp(β), exponential with $\beta = 1$, 2 or 4; Filter($-\log P$), filtering with $-\log P = 2$, 4 or 6; CG, CARDIoGRAM; SCZ, schizophrenia study; eGFRcrea, creatinine study.

The first four weighting schemes are: unweighted Bonferroni testing, where all weights equal unity; Spjøtvoll weights with parameters $\mu_i = (N_i/N_{0i})^{1/2}T_{0i}$; Bayes weights with dispersion $\phi = 0{\cdot}1$, 1 or 10; and exponential weights (Roeder et al., 2006), introduced in § 4·1, with tilt $\beta = 1$, 2 or 4.

The fifth and last weighting scheme is filtering, which selects the smallest p-values from the prior study and tests their hypotheses in the current study. We use three p-value thresholds, 10^{-2}, 10^{-4} and 10^{-6}. Rietveld et al. (2014) proposed a method for choosing the optimal p-value threshold for filtering, which requires the genotypic correlation between the two traits and the additive heritability of the current trait. For complex traits, these parameters are usually estimated with large uncertainty, and substantial domain expertise is needed to specify them.

We prune the significant single nucleotide polymorphisms for linkage disequilibrium using the DistiLD database (Palleja et al., 2012). Specifically, for each weighting scheme we select one locus from each linkage disequilibrium block that contains significant loci. Our data analysis pipeline is given in the Supplementary Material.

We compute a score $s_{m(p)d}$ for each weighting scheme m with parameters p, on each dataset d. This is defined as $+1$ if the weighting scheme increases the number of detections relative to unweighted testing, 0 if it leaves the number unchanged, and -1 otherwise. The score $s_{m(p)}$ of a weighting scheme m with parameters p is the sum of scores across datasets. The total s_m of the weighting scheme m is the sum of scores $s_{m(p)}$ across parameters.

6·3. *Results*

Table 1 shows the number of significant loci for each pair of studies and for each weighting scheme. We also present the results pruned for linkage disequilibrium, which act as a proxy for the number of independent loci found.

The results are somewhat inconclusive. In the positive control example, all weighting schemes except exponential weighting detect more loci than unweighted testing. Spjøtvoll weighting and filtering lead to the largest number of loci. In the blood lipids example, the methods generally detect fewer pruned loci, except for Bayes weights with $\phi = 10$. The methods can detect both a larger and a smaller number of unpruned loci, except in the case of Bayes weights, which uniformly increase the number of loci. For the eGFR creatinine example, exponential weights produce the best behaviour. We also see that the default $\phi = 1$ never performs worse than both unweighted testing and Spjøtvoll weights, and for the unpruned lipids example it is better.

If we allow tuning of parameters for the three weighting schemes that have such a parameter, Bayes weights show good performance: they are either first or second in all examples. This shows that our method is robust with respect to the choice of tuning parameter.

Finally, only Bayes weights with $\phi = 1$ or 10 have a positive score. The total score, summed across parameter settings, is also positive only for Bayes weights. Judging from these results, our method shows promise. However, from this analysis alone we cannot establish conclusively the relative merits of the methods. In future work it will be necessary to evaluate p-value weighting methods on more datasets.

Acknowledgement

Kristen Fortney and Stuart Kim are also affiliated with the Department of Genetics, Stanford University. This research was partially supported by the U.S. National Science Foundation and National Institutes of Health. We are grateful for the reviewers' constructive comments, which have helped to improve the paper.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of the theoretical results, software implementations in R and MATLAB, and code to reproduce the simulations and data analysis results.

References

Andreassen, O. A., Djurovic, S., Thompson, W. K., Schork, A. J., Kendler, K. S., O'Donovan, M. C., Rujescu, D., Werge, T., van de Bunt, M. & Morris, A. P. et al. (2013). Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am. J. Hum. Genet.* **92**, 197–209.

Benjamini, Y. & Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scand. J. Statist.* **24**, 407–18.

Bourgon, R., Gentleman, R. & Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proc. Nat. Acad. Sci.* **107**, 9546–51.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with Discussion). *J. R. Statist. Soc.* A **143**, 383–430.

Brooks-Wilson, A. R. (2013). Genetics of healthy aging and longevity. *Hum. Genet.* **132**, 1323–38.

Carlin, B. J. & Louis, T. A. (1985). Controlling error rates by using conditional expected power to select tumor sites. In *Proc. Biopharm. Sect., Am. Statist. Assoc.* Alexandria, Virginia: American Statistical Association, pp. 11–8.

Coronary Artery Disease Genetics Consortium (2011). A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nature Genet.* **43**, 339–44.

Darnell, G., Duong, D., Han, B. & Eskin, E. (2012). Incorporating prior information into association studies. *Bioinformatics* **28**, i147–53.

Eskin, E. (2008). Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome Res.* **18**, 653–60.

Genovese, C. R., Roeder, K. & Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika* **93**, 509–24.

Gui, J., Tosteson, T. D. & Borsuk, M. E. (2012). Weighted multiple testing procedures for genomic studies. *BioData Mining* **5**, article no. 4.

Hjelmborg, J., Iachine, I., Skytthe, A., Vaupel, J. W., McGue, M., Koskenvuo, M., Kaprio, J., Pedersen, N. L. & Christensen, K. (2006). Genetic influence on human lifespan and longevity. *Hum. Genet.* **119**, 312–21.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.

Köttgen, A., Pattaro, C., Böger, C. A., Fuchsberger, C., Olden, M., Glazer, N. L., Parsa, A., Gao, X., Yang, Q. & Smith, A. V. et al. (2010). New loci associated with kidney function and chronic kidney disease. *Nature Genet.* **42**, 376–84.

Louis, T. A. & Bailey, J. K. (1990). Controlling error rates using prior information and marginal totals to select tumor sites. *J. Statist. Plan. Infer.* **24**, 297–316.

Lynch, M. & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland: Sinauer Associates.

Palleja, A., Horn, H., Eliasson, S. & Jensen, L. J. (2012). DistiLD Database: Diseases and traits in linkage disequilibrium blocks. *Nucleic Acids Res.* **40**, D1036–40.

Peña, E. A., Habiger, J. D. & Wu, W. (2011). Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *Ann. Statist.* **39**, 556–83.

Rietveld, C. A., Esko, T., Davies, G., Pers, T. H., Turley, P., Benyamin, B., Chabris, C. F., Emilsson, V., Johnson, A. D. & Lee, J. J. et al. (2014). Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proc. Nat. Acad. Sci.* **111**, 13790–4.

Roeder, K. & Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statist. Sci.* **24**, 398–413.

Roeder, K., Bacanu, S.-A., Wasserman, L. & Devlin, B. (2006). Using linkage genome scans to improve power of association in genome scans. *Am. J. Hum. Genet.* **78**, 243–52.

Roquain, E. & Van De Wiel, M. A. (2009). Optimal weighting for false discovery rate control. *Electron. J. Statist.* **3**, 678–711.

Rubin, D., Dudoit, S. & Van der Laan, M. (2006). A method to increase the power of multiple testing procedures through sample splitting. *Statist. Applic. Genet. Molec. Biol.* **5**, 1–19.

Schizophrenia Psychiatric Genome-Wide Association Study Consortium (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature Genet.* **43**, 969–76.

Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., Stewart, A. F., Barbalic, M. & Gieger, C. et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genet.* **43**, 333–8.

Silverberg, D., Wexler, D., Blum, M., Schwartz, D. & Iaina, A. (2004). The association between congestive heart failure and chronic renal disease. *Curr. Opin. Nephrol. Hypertens.* **13**, 163–70.

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. (2013). Pleiotropy in complex traits: Challenges and strategies. *Nature Rev. Genet.* **14**, 483–95.

Spjøtvoll, E. (1972). On the optimality of some multiple comparison procedures. *Ann. Math. Statist.* **43**, 398–411.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I. & Willer, C. J. et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–13.

Westfall, P. H. & Krishen, A. (2001). Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *J. Statist. Plan. Infer.* **99**, 25–40.

Westfall, P. H. & Soper, K. A. (2001). Using priors to improve multiple animal carcinogenicity tests. *J. Am. Statist. Assoc.* **96**, 827–34.

Westfall, P. H., Krishen, A. & Young, S. S. (1998). Using prior information to allocate significance levels for multiple endpoints. *Statist. Med.* **17**, 2107–19.

Westfall, P. H., Kropf, S. & Finos, L. (2004). Weighted FWE-controlling methods in high-dimensional situations. In *Recent Developments in Multiple Comparison Procedures*, Y. Benjamini, F. Bretz and S. Sarkar, eds. Beachwood, Ohio: Institute of Mathematical Statistics, pp. 143–54.

[*Received April* 2015. *Revised August* 2015]