ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2016)
78, *Part* 4, *pp.* 781–804

Making the cut: improved ranking and selection for large-scale inference

Nicholas C. Henderson and Michael A. Newton

University of Wisconsin, Madison, USA

[Received May 2014. Final revision June 2015]

Summary. Identifying leading measurement units from a large collection is a common inference task in various domains of large-scale inference. Testing approaches, which measure evidence against a null hypothesis rather than effect magnitude, tend to overpopulate lists of leading units with those associated with low measurement error. By contrast, local maximum likelihood approaches tend to favour units with high measurement error. Available Bayesian and empirical Bayesian approaches rely on specialized loss functions that result in similar deficiencies. We describe and evaluate a generic empirical Bayesian ranking procedure that populates the list of top units in a way that maximizes the expected overlap between the true and reported top lists for all list sizes. The procedure relates unit-specific posterior upper tail probabilities with their empirical distribution to yield a ranking variable. It discounts high variance units less than popular non-maximum-likelihood methods and thus achieves improved operating characteristics in the models considered.

Keywords: Empirical Bayes; Posterior expected rank; *r*-value

1. Introduction

In all sorts of applications, data from a large number of measurement or inference units are processed to identify the most important units by some measure. This is certainly true in statistical genomics, where units might be genes, gene sets or single-nucleotide polymorphisms (SNPs), depending on the particular application, but it is also true more broadly. In agriculture investigators rank animals or plants by their breeding value (e.g. de los Campos *et al.* (2013)); performance evaluations in health and social sciences are common (e.g. Paddock and Louis (2011)). Typically, units are associated with unobserved real-valued parameters, and the importance of each unit is linked to the value of its parameter. A case that we consider is a genomewide association study examining risk factors for type 2 diabetes, in which the inference unit is the SNP, and the parameter of interest is a log-odds ratio measuring the effect on disease probability of SNP genotype (Morris *et al.*, 2012). A second case involves gene set enrichment among human genes that have been determined via ribonucleic acid (RNA) interference experiments to affect influenza virus replication (Hao *et al.*, 2013). Units here are sets of genes annotated to particular biological functions and parameters measure levels of enrichment. We develop two further examples to exercise the statistical issues: one from sports statistics (units are basketball players), and one from gene expression analysis (units are genes). If there had been no measurement error we would summarize each case by ranking units according to values of their parameters, focusing on the top of this list for further study. We consider here the inference task

Address for correspondence: Michael Newton, Department of Statistics, University of Wisconsin—Madison, 1300 University Avenue, Madison, WI 53706, USA.
E-mail: newton@stat.wisc.edu

to perform such ranking and selection from data. Whereas the emphasis of large-scale inference has been testing in relatively sparse settings (e.g. Efron (2010)), the present work addresses the inference task to rank order non-null parameters when the signal is relatively non-sparse.

A natural ranking is obtained by separately estimating unit-specific parameters, for instance by maximum likelihood applied locally to each unit. Since sampling fluctuations more easily put high variance units into the tails, units that are associated with relatively high standard error are overrepresented among the top units by this maximum likelihood estimate (MLE) ranking. Another commonly used procedure comes from large-scale hypothesis testing, where units are ranked by their p-value relative to a reference null hypothesis. Units that are associated with relatively low standard error are overrepresented among the top units by this ranking since both effect size and standard error affect testing power. Standard error in the type 2 diabetes case is affected by various factors including SNP allele frequency; set size affects standard error in the RNA interference case. When there is little variation between unit-specific standard errors, the different approaches give essentially the same assessment of the most important units. However, in many cases there is substantial variation in these standard errors, and quite different rankings can emerge.

For contemporary large-scale applications, the classical theory of ranking and selection leaves much to be desired. It addresses sampling probabilities like, 'under such-and-such a configuration of parameters and for sufficient amounts of data per unit the probability exceeds such-and-such that the true top j units are among the observed top k units' (e.g. Gibbons *et al.* (1979)). Although relevant to some tasks, these probabilities are difficult to work with and the resulting procedures are not often used in applied statistics. Theory is available on the sampling characteristics of empirical rankings (e.g. Xie *et al.* (2009) and Hall and Miller (2010)). Arguably, the thrust of methodological development for ranking and selection involves hierarchical modelling coupled with Bayes or empirical Bayes inference. Seminal contributions by Berger and Deeley (1988) and Laird and Louis (1989) helped to establish a framework that covers many contemporary applications and that has been elaborated in important ways (e.g. Shen and Louis (1998), Gelman and Price (1999), Wright *et al.* (2003), Lin *et al.* (2006), Brijs *et al.* (2007) and Noma *et al.* (2010)). We further elaborate this framework in an effort to provide a more effective generic method for large-scale inference, especially when large parameter units are in focus, when there are many units and when there is substantial variation in unit-specific standard errors.

Sampling artefacts of MLE and p-value ranking procedures, which were noted above, are well documented, but other approaches are also deficient. The insightful analysis of Gelman and Price (1999) illustrates the difficulties and confirms that the common practice of ranking by posterior expected value suffers from the same artefact as the p-value ranking, namely that units that are associated with small posterior standard deviation are overrepresented on lists of the top units. We find similar behaviour with the posterior expected rank method (Laird and Louis, 1989; Lin *et al.*, 2006) as well as available testing schemes. We introduce and investigate a procedure that aims to rank units in a way to maximize the expected overlap between the reported and the true top lists of units. Although not eliminating the sampling artefacts, the new method reduces their effects compared with other schemes. Our development starts in a special case wherein ranking procedures are formulated in terms of certain threshold functions (Section 2.1); using this formulation we characterize thresholds that maximize the expected overlap between the true and reported top lists (Sections 2.2 and 2.3), and we derive the associated ranking variable in terms of local posterior tail probabilities. The proposed *r-value* is generalized in Section 2.4 and investigated in relation to other procedures in Section 3. Computational issues are reviewed in Section 4, sampling performance is investigated in Section 5 and a short discussion follows. Examples are used throughout for demonstration, and proofs are

postponed until Appendix A. The methodology proposed and several data sets are deployed in the R package `rvalues`, which is available through the Comprehensive R Archive Network (`http://cran.r-project.org`).

2. Threshold functions and ranking variables

2.1. Continuous model

A variety of data structures are amenable to our proposed ranking or selection scheme, but the following structure has guided its initial development. Measurement or inference units are indexed by $i = 1, 2, \ldots, n$; data on unit i include the real-valued measurement X_i and information about its sampling variation. We assume that the sampling distribution of X_i has a known form that is indexed by an unknown real-valued parameter of interest θ_i together with a second quantity affecting variance. In this section we assume that $\sigma_i^2 = \mathrm{var}(X_i)$ is known for each unit. Basically, the inference task is to report units having large values of θ_i, while accounting for the fact that variances σ_i^2 may fluctuate substantially between inference units. We adopt an empirical Bayes perspective and treat $\{(\theta_i, \sigma_i^2)\}$ as draws from a population of parameters, and we are motivated by data analysis considerations to suppose initially that θ_i and σ_i^2 are independent in this population, say with densities $f(\theta)$ and $g(\sigma^2)$. The independence assumption is helpful for understanding artefacts of various ranking methods, but it is not essential to the methodology. The empirical Bayesian uses the full data set to estimate the *prior* distributions $f(\theta)$ and $g(\sigma^2)$. Initially we ignore the estimation error at this level and focus on ranking units within the estimated population, though we take up the issue in Section 5 via simulation and asymptotic analysis.

Relative to a single unit i, X_i might be the maximum likelihood estimator of θ_i, and σ_i that estimator's standard error. The independence assumption may be reasonable if care has been taken in this local analysis, for example, by variance stabilizing transformation. Typically, the variance σ_i^2 is estimated rather than known exactly; we study this and extensions to other data structures in Section 2.4. We consider first a continuous model, involving prior distributions and sampling distributions all having densities with respect to Lebesgue measure. The canonical sampling model within this class has $X_i|\theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2)$.

We make some headway by associating each ranking or selection procedure with a family $\mathcal{T}$ of threshold functions $\mathcal{T} = \{t_\alpha : \alpha \in (0, 1)\}$. Each t_α is a function $t_\alpha(\sigma^2)$ having the interpretation that unit i is reported to be in the top α fraction of units if and only if $X_i \geqslant t_\alpha(\sigma_i^2)$. This inter-

Table 1. Threshold functions associated with various ranking criteria, normal–normal model

Criterion	*Ranking variable*	*Threshold function* $t_\alpha(\sigma^2)$
MLE	X_i	u_α
p-value $H_0 : \theta_i = 0$	X_i/σ_i	$u_\alpha \sigma$
p-value $H_0 : \theta_i = c$	$(X_i - c)/\sigma_i$	$c + u_\alpha \sigma$
PM	$X_i/(\sigma_i^2 + 1)$	$u_\alpha(\sigma^2 + 1)$
PER	$P(\theta_i \leqslant \theta | X_i, \sigma_i^2)$	$u_\alpha \sqrt{\{(\sigma^2 + 1)(2\sigma^2 + 1)\}}$
Bayes factor	$\mathbf{1}(X_i > 0) \dfrac{P(X_i | \sigma_i^2, \theta_i \neq 0)}{P(X_i | \sigma_i^2, \theta_i = 0)}$	$\sqrt{(\sigma^2(\sigma^2 + 1)[u_\alpha + \log\{(\sigma^2 + 1)/\sigma^2\}])}$
Maximal agreement	r-value	$\theta_\alpha(\sigma^2 + 1) - u_\alpha \sqrt{\{\sigma^2(\sigma^2 + 1)\}}$

pretation is supported by the *size constraint*, namely that, marginally to all parameters and data,

$$P\{X_i \geqslant t_\alpha(\sigma_i^2)\} = \alpha \qquad \text{for all } \alpha \in (0,1). \tag{1}$$

Table 1 reports threshold functions associated with a variety of ranking methods in the normal observation model, and under the extra condition that the prior $f(\theta)$ is $N(\mu, \tau^2)$. Table 1 encodes the special case $(\mu, \tau^2) = (0, 1)$; the general thresholds are derived from this case by the transformation $\mu + \tau\, t_\alpha(\sigma^2/\tau^2)$. Note that each threshold function involves an α-specific value u_α which guarantees the size constraint; these values are different for different ranking methods (rows of Table 1). Fig. 1 illustrates four of these families in the type 2 diabetes case-study. Notionally, the linear ranking of units is obtained by sweeping through the family $\mathcal{T}$, beginning with the smallest α at the top of the graph. Clearly, distinct families of threshold functions can produce distinct rankings of the units, with the family's shape revealing how it trades off observed signal X_i with measurement variance σ_i^2 to prioritize the leading units.

Some comments on the threshold functions in Table 1 are warranted (see also the on-line supplementary material document). Under squared error loss, the Bayes estimate of the rank

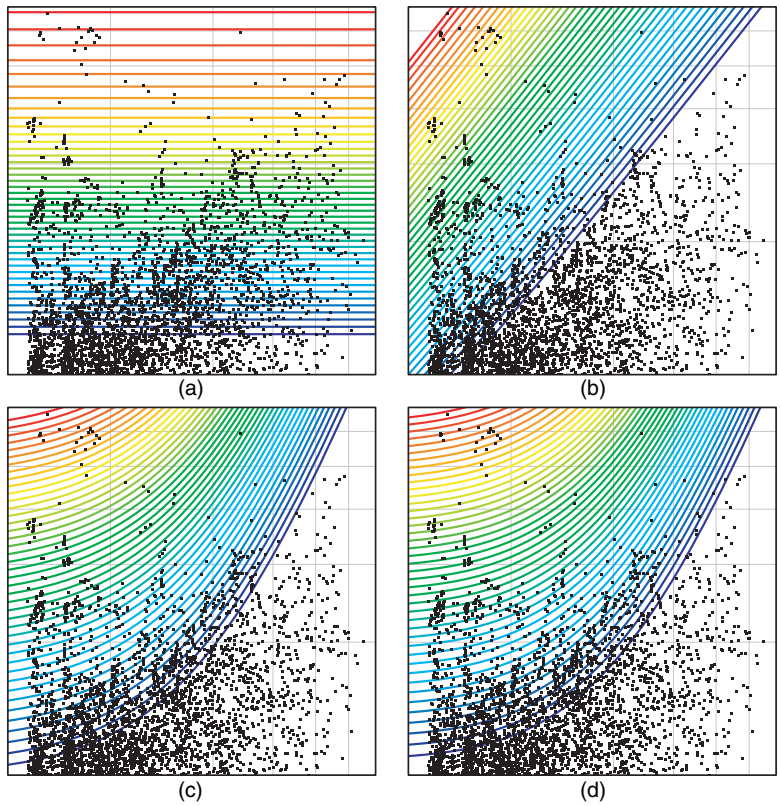

Fig. 1. Threshold functions (a) MLE, (b) *p*-value, (c) PM and (d) maximal agreement, type 2 diabetes example: axes are common to all panels, with the vertical axis the log-odds ratio for association between SNPs (■) and type 2 diabetes and with the horizontal axis the standard error estimates, with further details in Fig. S1 in the on-line supplementary material; calculations use an inverse gamma model for σ^2; 42 threshold functions are shown ranging in α-values from a small positive value (——) just including the first data point up to $\alpha = 0.10$ (——) (most SNPs are truncated by the plot; also the grid is uniform on the scale of $\log_2\{-\log_2(\alpha)\}$); units associated with a smaller α (i.e. more red) are ranked more highly by the given ranking method; two units landing on the same curve would be ranked in the same position

of parameter θ_i among those in play is the conditional expected rank given the data (Laird and Louis, 1989; Lin *et al.*, 2006). This posterior expected rank is usually expressed as a sum, involving indicator comparisons between θ_i and the other parameters, and it becomes $P(\theta_i \leqslant \theta | X_i, \sigma_i^2)$ when normalized and considered in the limit for increasing numbers of units (ranking from the top). Here θ is the independently drawn parameter of a generic additional unit, which emerges in the large-scale limit to replace the collection of all other θ_js with which θ_i is compared. In the normal–normal model, ranking by posterior expected rank is qualitatively similar to ranking by posterior mean PM $= E(\theta_i | X_i, \sigma_i^2)$; both favour small variance units. Several hypothesis-testing-based methods are also shown in Table 1. Testing against some benchmark null (rather than the no-effect null) hypothesis has some benefits in practice (e.g. McCarthy and Smyth (2009)). As we emphasize large positive θ_i, we report p-values that are associated with one-sided tests. Finally, the Bayes factor BF entry aims to mimic the ranking (from the top) method that is associated with Bayes factors for the test of $H_0 : \theta_i = 0$ *versus* $H_A : \theta_i \neq 0$ (e.g. Kass and Raftery (1995)). The mapping of a ranking method to a family of threshold functions is useful for comparative analyses, as we investigate next.

2.2. Thresholds via direct optimization

Table 1 and Fig. 1 introduce a family $\mathcal{T}^* = \{t_\alpha^*\}$ that is optimal in the continuous model in the sense that for all $\alpha \in (0, 1)$

$$P\{X_i \geqslant t_\alpha^*(\sigma_i^2), \theta_i \geqslant \theta_\alpha\} \geqslant P\{X_i \geqslant t_\alpha(\sigma_i^2), \theta_i \geqslant \theta_\alpha\} \tag{2}$$

for any other family $\mathcal{T} = \{t_\alpha\}$ which also satisfies the size constraint (1). Here θ_α is the α upper quantile of the prior, i.e. $P(\theta_i \geqslant \theta_\alpha) = \alpha$. In other words, $\mathcal{T}^*$ maximizes *agreement*: the joint probability that unit i is placed in the top α fraction and its driving parameter θ_i is in the top α fraction of the population, for all α. We emphasize that the probabilities in inequality (2) cover the joint distribution of $(X_i, \sigma_i^2, \theta_i)$, which respects both the sampling distribution of data local to unit i and the fluctuations of unit-specific parameters. A calculus-of-variations argument provides direct optimization of the joint probability in inequality (2), subject to the size constraint, model regularity and smoothness of the threshold functions.

Theorem 1. In the continuous model, a necessary condition for the function t_α^* to be optimal as in inequality (2), within the class of continuously differentiable threshold functions, is that it satisfies

$$P\{\theta_i \geqslant \theta_\alpha | X_i = t_\alpha^*(\sigma^2), \sigma_i^2 = \sigma^2\} = c_\alpha \qquad \text{for all } \sigma^2. \tag{3}$$

Thus, all observations coincident with the graph of a given optimal threshold curve have a common posterior probability c_α that their unit-specific parameters exceed the quantile θ_α that is associated with that curve. In the normal model for X_i and the normal prior $f(\theta)$, the optimal threshold function (Fig. 1(d)) is readily extracted from expression (3). Working on a standardized scale without loss of generality ($\mu = 0$ and $\tau^2 = 1$), the local posterior for θ_i is normal with mean $X_i/(\sigma_i^2 + 1)$ and variance $\sigma_i^2/(\sigma_i^2 + 1)$. Thus,

$$t_\alpha^*(\sigma^2) = \theta_\alpha(\sigma^2 + 1) - u_\alpha \sqrt{\{\sigma^2(\sigma^2 + 1)\}}, \tag{4}$$

where $\theta_\alpha = \Phi^{-1}(1 - \alpha)$ and u_α is determined by the size constraint (1). Indeed u_α is affected by the distribution $g(\sigma^2)$, since it is defined implicitly by

$$1 - \alpha = \int_0^\infty \Phi\{\theta_\alpha \sqrt{(\sigma^2 + 1)} - u_\alpha \sigma\}\, g(\sigma^2) \mathrm{d}\sigma^2 \tag{5}$$

where Φ is the standard normal cumulative distribution function.

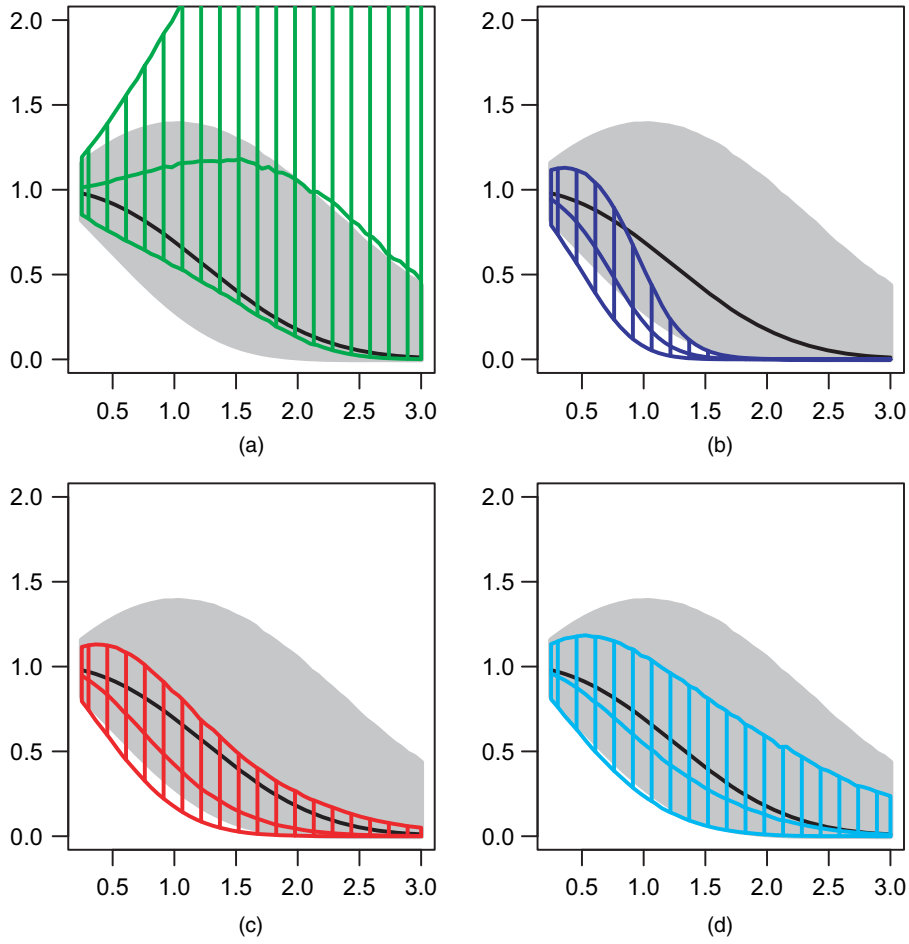

Fig. 2. Conditional distribution (median, interquartile range) of unit-specific variance σ_i^2 given selection of the unit in the top $\alpha = 0.1$ fraction by various methods (coloured bands) compared with the marginal gamma distribution (black or grey) for different amounts of variation in σ_i^2 ($E(\sigma^2) = 1$; coefficient of variation on the horizontal axis; based on simulation using 10^7 units per case): (a) MLE; (b) *p*-value; (c) PM; (d) maximal agreement

Curiously, the optimal thresholds *kick up* as σ^2 approaches 0. The resolution and range of Fig. 1 do not reveal this phenomenon so clearly in the type 2 diabetes example, but it is apparent from equation (4) that the derivative of t_α^* with respect to σ^2 becomes increasingly negative as σ^2 approaches 0 (when $u_\alpha > 0$). Neither the *p*-value thresholds nor those based on posterior mean or posterior expected rank have this characteristic; indeed, by kicking up for small σ^2 the maximal agreement thresholds are less prone to the overranking of small variance units.

Fig. 2 illustrates sampling properties of top-listed units obtained by various threshold schemes, including the optimal threshold (4), using the normal–normal model, $\alpha = 0.1$, a sequence of gamma distributions g for the variance σ_i^2 and independence between θ_i and σ_i^2. The difference between different methods becomes more pronounced as we increase variation in the distribution of the variances; in this simulation all cases involve $E(\sigma_i^2) = 1$, but the shape parameters vary to increase the coefficient of variation. The degree to which the conditional distribution of

σ_i^2 given placement on the top α list (the coloured bars) differs from the marginal distribution of σ_i^2 in the system (grey) measures the extent of sampling artefacts by that method. The example recapitulates sampling artefacts of the local MLE, the p-value and the posterior mean. For example, the top lists by MLE are enriched for high variance units. Fig. 2 also shows that this artefact is substantially reduced when we select by equation (4).

2.3. Posterior tail probabilities and ranking variables

Except in stylized models we cannot solve equation (3) to identify optimal thresholds for ranking. Insight into their structure comes by further examining their relationship to local posterior tail probabilities: $V_\alpha(X_i, \sigma_i^2) = P(\theta_i \geqslant \theta_\alpha | X_i, \sigma_i^2)$.

Theorem 2. Suppose that for $\alpha \in (0, 1)$ there exists λ_α such that

$$P\{V_\alpha(X_i, \sigma_i^2) \geqslant \lambda_\alpha\} = \alpha, \tag{6}$$

and furthermore that $V_\alpha(x, \sigma^2)$ is right continuous and non-decreasing in x for fixed α and σ^2. Then the family of thresholds $t_\alpha^*(\sigma^2) = \inf\{x : V_\alpha(x, \sigma^2) \geqslant \lambda_\alpha\}$ satisfies the size constraint (1) and is optimal in the sense of inequality (2).

These conditions concern V_α and its distribution. They are satisfied in the normal–normal model; there,

$$V_\alpha(x, \sigma^2) = 1 - \Phi\left\{\sqrt{\left(\frac{\sigma^2+1}{\sigma^2}\right)}\left(\theta_\alpha - \frac{x}{\sigma^2+1}\right)\right\}$$

and $\lambda_\alpha = 1 - \Phi(u_\alpha)$, for u_α is as in equation (5). The conditions are also satisfied in other instances of the continuous model of Section 2.1, as well as in other settings. For example, if σ_i^2 is an estimated variance, then a Student t sampling model might replace the normal sampling model conditional on σ_i^2 and θ_i. See the on-line supplementary material for this and other examples. Note that the optimal threshold $t_\alpha^*(\sigma^2)$ in theorem 2 simplifies further if $V_\alpha(x, \sigma^2)$ is continuous and strictly increasing in x for each α and σ^2. Then $t_\alpha^*(\sigma^2) = V_\alpha^{-1}(\lambda_\alpha, \sigma^2)$, with the inverse referring to the first (i.e. x) argument.

A family of threshold functions is a device to think about converting observations into rankings (i.e. by sweeping through the family). Indeed, the index α that is associated with the threshold curve on which data point (X_i, σ_i^2) lands is a ranking variable; its computation amounts to solving the inversion $X_i = t_\alpha(\sigma_i^2)$ for α. Exact inversion is possible as long as the threshold curves for different α-values do not cross, i.e. if there are no values $\alpha_1 < \alpha_2$, σ^2 for which $t_{\alpha_1}(\sigma^2) = t_{\alpha_2}(\sigma^2)$. Approximate inversion is always possible via $\inf\{\alpha : X_i \geqslant t_\alpha(\sigma^2)\}$.

Theorem 3. Suppose that threshold functions $t_\alpha(\sigma^2)$ are differentiable in α for each σ^2. No functions in the family cross as long as $\partial t_\alpha(\sigma^2)/\partial\alpha < 0$ for every $\alpha \in (0, 1)$. Further, the optimal thresholds in the normal–normal model do not cross.

This confirms more generally what we see empirically for a few cases in Fig. 1 and Table 1: the optimal thresholds do not cross under the conditions of theorem 3, and they conform to our intuition about how ranking procedures might be constructed from threshold functions.

We introduce a special ranking variable that inverts the optimal threshold. For the ith unit, we define the r-value

$$r(X_i, \sigma_i^2) = \inf\{\alpha : V_\alpha(X_i, \sigma_i^2) \geqslant \lambda_\alpha\}. \tag{7}$$

Essentially, unit i is placed by its r-value at position α (a relative rank, measured from the top) if, when ranking the units by $V_\alpha(X_i, \sigma_i^2)$, it also happens to land at position α. Further, the top

α fraction of units by r-value has higher overlap with the true top α fraction of units than could be obtained by any other ranking procedure, in the sense of inequality (2).

It is worth recognizing that these findings go beyond what has been reported about the use of the conditional tail probability $V_\alpha(X_i, \sigma_i^2)$ to rank units. Classical theory on optimal selection establishes the role of this conditional tail probability in maximizing an exceedance probability within the selected sample (e.g. Lehmann (1986), pages 117–118). Also, the conditional tail probability has been used for ranking (e.g. Normand *et al.* (1997) and Niemi (2010)) and is closely related to a Bayes optimal ranking under a certain loss function (Lin *et al.*, 2006). A critical difference with the ranking proposed is in the role of the index α. Conceptually, we imagine ranking the units by $V_\alpha(X_i, \sigma_i^2)$ separately for all possible indices α (not just a prespecified index); then the r-value for unit i is the smallest index α such that unit i is placed in the top α fraction by that ranking. By aiming to maximize agreement at all list sizes, the method proposed does not require a prespecified exceedance level to generate its ranking.

2.4. More generality

The r-value construction makes sense in various elaborations of the model from Section 2.1. We retain univariate parameters of interest $\{\theta_i\}$ varying according to a distribution F, but we allow data D_i on each unit to take more general forms than the (X_i, σ_i^2) pair structure. We also retain the assumption of mutual independence between units, though extensions could be developed in cases where posterior computation is feasible. In seeking units with largest θ_i, the critical quantity is the local exceedance probability, $V_\alpha(D_i) = P(\theta_i \geqslant \theta_\alpha | D_i)$, for $\alpha \in (0, 1)$ and for upper quantiles θ_α of the marginal distribution F, i.e. $\theta_\alpha = F^{-1}(1-\alpha)$. Induced by the marginal distribution of D_i, the tail probability $V_\alpha(D_i)$ has cumulative distribution function $H_\alpha(v)$, and from it we obtain the upper quantile: $\lambda_\alpha = H_\alpha^{-1}(1-\alpha)$. Then, by analogy with definition (7), the r-value is defined: $r(D_i) = \inf\{\alpha : V_\alpha(D_i) \geqslant \lambda_\alpha\}$.

Fig. 3 compares r-value rankings with three other methods in the RNA interference example. Here, $D_i = (m_i, y_i)$ holds binomial information (set size m_i and number y_i of genes in set i that were identified by RNA interference). The target parameters θ_i are treated as draws from a beta(a, b) distribution, with shape parameters estimated by marginal maximum likelihood, and the conditional tail probability $V_\alpha(D_i)$ becomes the probability that a beta$(\hat{a} + y_i, \hat{b} + m_i - y_i)$ variable exceeds θ_α. r-value computation (see Section 4) requires the sampling distribution of these tail probabilities, which we approximated by using the data from all 5719 sets under study. The methods compared in Fig. 3 agree to some extent on the ranking of the most interesting sets, but systematic differences are apparent. Ranking by y_i/m_i overranks small sets; ranking by p-value overranks large sets; and ranking by posterior mean $(y_i + \hat{a})/(m_i + \hat{a} + \hat{b})$ also overranks large sets, though to a lesser degree, all compared with the r-value ranking.

Sports enthusiasts routinely rank players. To explore r-value ranking in this context, we deploy the same beta–binomial model as used in the RNA interference example and use it to describe free-throw statistics of professional basketball players (e.g. Richey and Zorn (2005)). During the 2013–2014 regular season of the National Basketball Association (NBA), 461 players attempted at least one free throw (Entertainment and Sports Programming Network, 2014). In total these players attempted 58 029 free throws and were successful 43 870 times, for a marginal free-throw percentage of 75.6%. A basic problem in rating players by individual free-throw percentage FTP $= y_i/m_i$ is that the numbers $\{m_i\}$ of free-throw attempts vary substantially between players; in retaining all active players, those with highest y_i/m_i are among those with smallest m_i. For instance, 13 of the 461 NBA players had perfect free-throw records in 2013–2014; they had a median number of four attempts, compared with the league median of 82

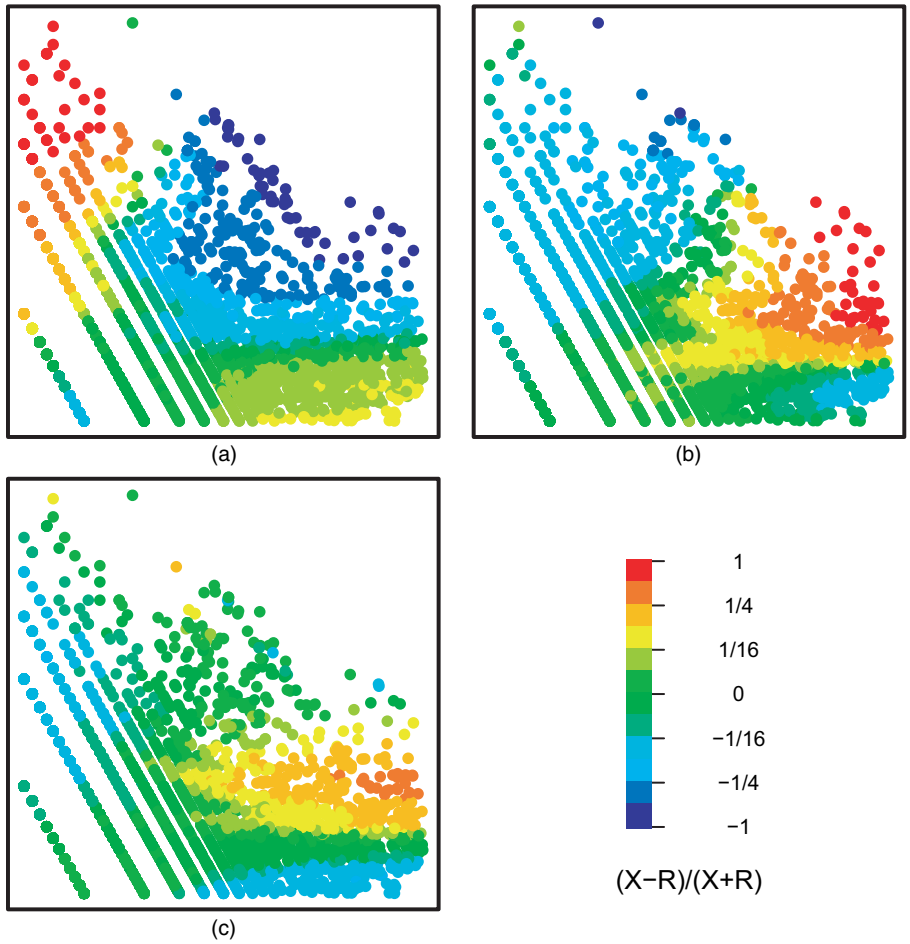

Fig. 3. Ranking via various methods compared with *r*-value ranking, RNA interference example (the data and axes are common to all panels, with further details in Fig. S2 in the on-line supplementary material; briefly the horizontal axis is set size (on a log-scale) and the vertical axis is gene set enrichment; each set (dot) is coloured by $(X - R)/(X + R)$ where X is the rank (from the top) of the set by the method being compared, and R is the rank by *r*-value): (a) MLE; (b) *p*-value; (c) PM

attempts. Various threshold schemes have been adopted by rating agencies; these restrict ranking to players reaching a minimum number of attempts or a minimum number of makes. At the Entertainment and Sports Programming Network, a *qualified* player this last season needed $y_i \geqslant 125$. Thresholding rules have a practical appeal but they can suppress athletic performances that otherwise are exceptional and worth reporting. For instance, Ray Allen's 105 makes in 116 attempts is exceptionally good by many standards (Table 2). The context provided by the NBA example offers further insights. For one thing, there is broad agreement between PM-ranking and *r*-value ranking, though where there is disagreement PM favours players having more attempts m_i. Related to this is the fact that, though it discounts players with very small m_i, the *r*-value shrinks less than PM and is more in accordance with the FTP ranking; for example, the *r*-value ranks the qualified players in Table 2 the same as FTP, in contrast with PM.

As an empirical validation of the *r*-value ranking we applied it to mid-season NBA data (up to the end of December 2013) and then measured its performance conditionally on complete-

Table 2. Leading free-throw shooters, 2013–2014 regular season of the NBA†

Player i	y_i	m_i	*FTP*	*PM*	*RV*	*QR*	*MLER*	*PMR*	*RVR*
Brian Roberts	125	133	0.940	0.913	0.002	1	17	1	1
Ryan Anderson	59	62	0.952	0.898	0.003		15	2	2
Danny Granger	63	67	0.940	0.893	0.005		16	3	3
Kyle Korver	87	94	0.926	0.892	0.008		19	4	4
Mike Harris	26	27	0.963	0.866	0.010		14	15	5
J. J. Redick	97	106	0.915	0.886	0.011		22	6	6
Ray Allen	105	116	0.905	0.880	0.016		25	8	7
Mike Muscala	14	14	1.000	0.844	0.017		7	34	8
Dirk Nowitzki	338	376	0.899	0.891	0.018	2	30	5	9
Trey Burke	102	113	0.903	0.877	0.018		28	9	10
Reggie Jackson	158	177	0.893	0.877	0.024	3	32	11	11
Kevin Martin	303	340	0.891	0.882	0.025	4	33	7	12
Gary Neal	94	105	0.895	0.869	0.025		31	14	13
D. J. Augustin	201	227	0.885	0.873	0.031	5	38	12	14
Stephen Curry	308	348	0.885	0.877	0.031	6	39	10	15
Patty Mills	73	82	0.890	0.860	0.032		34	19	16
Courtney Lee	99	112	0.884	0.861	0.035		40	18	17
Steve Nash	22	24	0.917	0.834	0.039		20.5	44	18
Greivis Vasquez	95	108	0.880	0.857	0.040		41	22	19
Robbie Hummel	15	16	0.938	0.825	0.043		18	55	20
Mo Williams	78	89	0.876	0.850	0.046		42	24	21
Kevin Durant	703	805	0.873	0.870	0.048	7	45	13	22
Aaron Brooks	83	95	0.874	0.850	0.049		44	26	23
Damian Lillard	371	426	0.871	0.865	0.050	8	47	16	24
Nando de Colo	31	35	0.886	0.831	0.057		37	48	25

†From $n = 461$ players who attempted at least one free throw, shown are the top 25 players as inferred by r-value. Data D_i on player i include the number of made free throws y_i and the number of attempts m_i. Other columns indicate the free-throw percentage FTP $= y_i/m_i$, which is the MLE of the underlying ability θ_i; posterior mean $E(\theta_i|D_i)$, r-value $\inf\{\alpha \geqslant 1/n : P(\theta_i \geqslant \theta_\alpha|D_i) \geqslant \lambda_\alpha\}$; qualified rank QR, which is the rank of FTP among players for whom $y_i \geqslant 125$; and ranks associated with the MLE, posterior mean and r-value.

season data. Comparing Table S2 (in the on-line supplementary material) with Table 2, we see some interesting features. For example, Brian Roberts, who finished the season with the highest FTP among qualified players, did not miss in 2013; the r-value placed him second mid-season, even though he had only $m_i = 18$ attempts, whereas PM ranked him 12th. Investigating more fully, we repeatedly simulated $\{\theta_i\}$-vectors conditionally on end-of-season data and averaged a similarity score:

$$\frac{1}{t}\sum_{i=1}^{t} \mathbf{1}\{\text{rank}(\theta_i) \leqslant t\}\mathbf{1}(\widehat{\text{rank}}_i \leqslant t),$$

finding improvements over FTP and PM in assessing the best free-throw shooters (Fig. S3 in the supplementary material). Here $\widehat{\text{rank}}_i$ is the player's estimated rank according to mid-season data and $\text{rank}(\theta_i)$ is his unknown true rank.

r-values may be computed in all sorts of hierarchical modelling efforts, including semiparametric models and cases where Markov chain Monte Carlo sampling is used to approximate the marginal posterior distribution of each θ_i given available data. Fig. S9 (in the supplementary material) compares the r-value ranking with other rankings in an example from gene expression analysis, where evidence suggested that the expression of a large fraction of the human genome was associated with the status of a certain viral infection (Pyeon *et al.*, 2007). A multilevel model

involving both null and non-null genes as well as t-distributed non-null effects θ_i exhibited good fit to the data but did not admit a closed form for $V_\alpha(D_i)$. r-values, computed by using Markov chain Monte Carlo output, again reveal systematic ranking differences from other approaches.

Multilevel models drive statistical inference and software in a variety of genomic domains, e.g. `limma` (Smyth, 2004), `EBarrays` (Kendziorski *et al.*, 2003) and `EBSeq` (Leng *et al.*, 2013), among others. Since these models happen to specify distributional forms for parameters of interest, the associated code could be augmented to compute posterior tail probabilities $V_\alpha(D_i)$ and thus r-values for ranking. The `limma` system utilizes a conjugate normal, inverse gamma model, and so $V_\alpha(D_i)$ involves the tail probability of a non-central t-distribution. The `EBSeq` system entails a conjugate beta, negative binomial model, and so $V_\alpha(D_i)$ for differential expression involves tail probabilities in a certain ratio distribution (Coelho and Mexia, 2007). One expects the benefits of r-value computation to show especially in cases involving many non-null units and relatively high variation between units in their variance parameters (e.g. sequence read depth). The data structure that is envisioned for r-value computation involves many exchangeable units, with real-valued parameters driving the conditional distribution of data on each unit. Other structures, such as from large-scale regression, may be amenable to the proposed ranking method if marginal posterior distributions for each regression coefficient could be derived.

3. Connections

3.1. Connection to Bayes rule

The proposed r-values are not Bayes rules in the usual sense; however, there is a connection to Bayesian inference if we allow both a continuum of loss functions and a distributional constraint on the reported unit-specific (relative) ranks. To see this connection, we introduce a collection of loss functions

$$L_\alpha(a, \theta_i) = 1 - \mathbf{1}(a \leqslant \alpha, \theta_i \geqslant \theta_\alpha)$$

where action a is a relative rank value in $(0, 1)$, $\alpha \in (0, 1)$ indexes the collection and again $\theta_\alpha = F^{-1}(1-\alpha)$ is a quantile in the population of interest. Specifically, no α-loss occurs if the inferred relative rank a and the actual relative rank $1 - F(\theta_i)$ both are less than α. The marginal (preposterior) Bayes risk of rule $\delta(D_i)$ is

$$\text{risk}_\alpha = 1 - P\{\delta(D_i) \leqslant \alpha, \theta \geqslant \theta_\alpha\}, \tag{8}$$

which is 1 minus the agreement (2). In the absence of other considerations, the Bayes rule for loss L_α degenerates to $\delta(D_i) = 0$. Degeneration is avoided if we enforce on the reported rank the additional structure that it shares with the true relative rank $1 - F(\theta_i)$ the property of being uniformly distributed over the population of units. Such a constrained Bayes rule then minimizes the modified objective function: $\text{risk}_\alpha + \gamma_\alpha P\{\delta(D_i) \leqslant \alpha\}$, where γ_α is chosen to enforce the (marginal) size constraint $P\{\delta(D_i) \leqslant \alpha\} = \alpha$.

The constrained Bayes rule is computed conditionally, per observed D_i, by minimizing the constraint-modified posterior expected loss PEL:

$$\begin{aligned} \text{PEL}_\alpha &= 1 - P\{\delta(D_i) \leqslant \alpha, \theta_i \geqslant \theta_\alpha | D_i\} + \gamma_\alpha \mathbf{1}\{\delta(D_i) \leqslant \alpha\} \\ &= \begin{cases} 1 - V_\alpha(D_i) + \gamma_\alpha & \text{if } \delta(D_i) \leqslant \alpha, \\ 1 & \text{if } \delta(D_i) > \alpha \end{cases} \end{aligned} \tag{9}$$

where $V_\alpha(D_i)$ is the upper posterior probability $P(\theta_i \geqslant \theta_\alpha | D_i)$ appearing in Section 2.

Curiously, a rule minimizing PEL_α is not uniquely determined at a single α, since minimization in equation (9) requires only that

$$\delta(D_i) \leqslant \alpha \Leftrightarrow V_\alpha(D_i) \geqslant \gamma_\alpha. \tag{10}$$

However, taking all losses together does fix a procedure. To see this, let $g(\alpha|D_i) = V_\alpha(D_i) - \gamma_\alpha$, and further assume that g is continuous in α. If $g(\alpha|D_i)$ has only one root in $(0,1)$, then the procedure $\delta^*(D_i) = \inf\{\alpha : V_\alpha(D_i) \geqslant \gamma_\alpha\}$ is a Bayes rule for any choice of L_α, even though δ^* does not depend on any specific choice of α. This is because $\delta^*(D_i) \leqslant \alpha$ for all α such that $g(\alpha|D_i) \geqslant 0$, and $\delta^*(D_i) > \alpha$ for all α such that $g(\alpha|D_i) < 0$. If $g(\alpha|D_i)$ does contain multiple roots (at least over a range of D_i that has positive probability), there will not be a procedure (i.e. a procedure which does not depend on α) which is a Bayes rule for any choice of L_α. This is because it will not be possible to construct a rule δ that satisfies requirement (10) for all values of $\alpha \in (0,1)$. The thresholds γ_α in expression (10) are determined by the uniformity constraint, and we have $\gamma_\alpha = H_\alpha^{-1}(1-\alpha)$, where H_α is the marginal distribution of $V_\alpha(D_i)$, counting all sources of variation, and so $\gamma_\alpha = \lambda_\alpha$ from the previous section. In other words, the procedure that is obtained by this constrained, multiloss Bayes calculation is equivalent to the *r*-value that was introduced in Section 2.

Among the more popular loss-based ranking procedures is one via posterior expected rank PER (e.g. Laird and Louis (1989) and Noma *et al.* (2010)). Unit i's value becomes $\mathrm{PER}_i = P(\theta_i \leqslant \theta|D_i)$ after normalizing by the number of units and taking the large-scale limit. We find in numerical experiments that PER-ranking is relatively close to the ranking by PM, and in these experiments we use $\mathrm{PER}_i = 1 - \int_0^1 V_\alpha(D_i)\,\mathrm{d}\alpha$, which can be established readily by using a transformation-of-variables argument.

3.2. Beyond ps and qs

In testing a single hypothesis H_0, the sample space may be structured as a nested sequence of subsets, $\{\Gamma_\alpha : \alpha \in (0,1)\}$, say, such that rejection of a size α test is equivalent to data D landing in set (i.e. rejection region) Γ_α. Then, the *p*-value of the test is $p(D) = \inf\{\alpha : D \in \Gamma_\alpha\}$. Storey (2003) extended this idea to multiple testing and the positive false discovery rate with the introduction of the *q*-value. Specifically, with another nested sequence $\{\tilde{\Gamma}_\alpha : \alpha \in (0,1)\}$ indexed such that $P(H_0|D \in \tilde{\Gamma}_\alpha) = \alpha$, the *q*-value is $q(D) = \inf\{\alpha : D \in \tilde{\Gamma}_\alpha\}$. Where *p*-values refer to the distribution of D on H_0, and *q*-values the conditional probability of H_0 given sample information, the proposed *r*-values refer to marginal probability over both unit-specific data and unit-specific parameters. The size constraint (1) corresponds to another sequence of subsets, $\{\check{\Gamma}_\alpha\}$, say, for which the marginal constraint holds: $P(D \in \check{\Gamma}_\alpha) = \alpha$. Analogously, the *r*-value is $r(D) = \inf\{\alpha : D \in \check{\Gamma}_\alpha\}$. In principle an *r*-value could be defined for any indexed ranking method, though we have reserved the definition for that method which maximizes agreement (2). Other connections to hypothesis testing are discussed in the on-line supplementary material.

4. Computation

In Section 2.2 we focused on the model involving normality for both the measurement X_i and the latent parameter θ_i. The *r*-value is obtained by inverting equation (4) to solve for r:

$$X_i = (\sigma_i^2 + 1)\,\Phi^{-1}(1-r) - u_r\sqrt{\{\sigma_i^2(\sigma_i^2+1)\}} \tag{11}$$

where u_r, defined through the size constraint (5), is readily computed numerically.

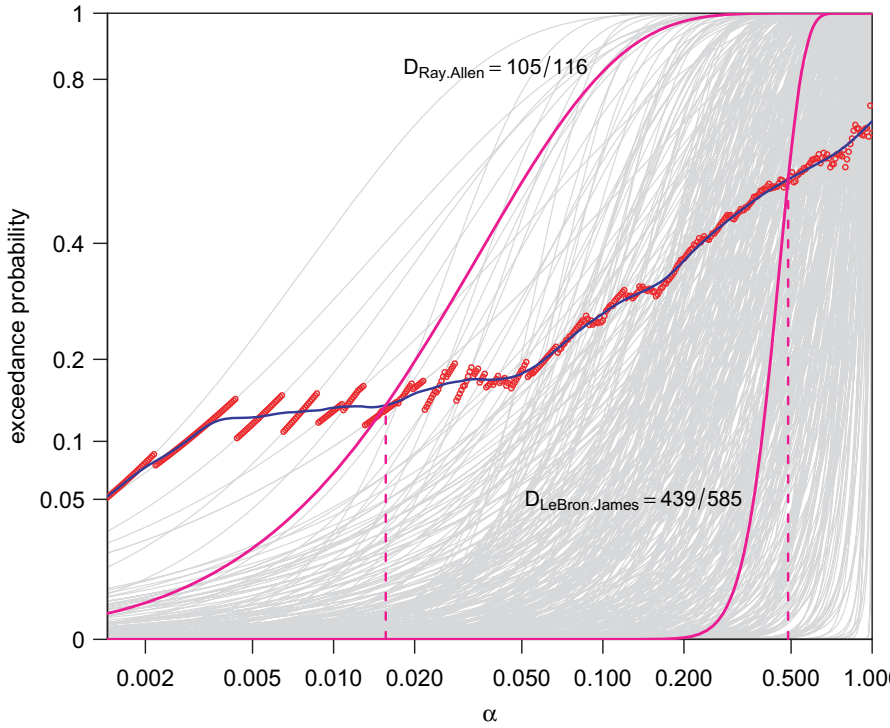

Fig. 4. Computational details, NBA example: ——, $P(\theta_i \geqslant \theta_\alpha | D_i)$; ——, two examples; ○, empirical quantile; ——, $\hat{\lambda}_\alpha$; ⋮, r-value

Alternatively, a generic approach to computing r-values starts with a finite grid $\{\alpha_j\}$ in $(0, 1)$, at which we compute the posterior tail probabilities $v_{i,j} = V_{\alpha_j}(D_i)$ for all units i (or approximations, e.g. by Markov chain Monte Carlo sampling). The grid need not be uniform; we enrich coverage near 0 in our implementation. The jth column of the matrix $\{v_{i,j}\}$ holds a sample from the marginal distribution for which λ_{α_j} is the $(1-\alpha_j)$-quantile. Marching through j allows us to assemble a discrete (in α), empirical (over units) quantile function, which we convert to a function $\hat{\lambda}_\alpha$ first by possibly smoothing to mitigate sampling effects and then by interpolating to α-values beyond the initial grid. Then for each unit i we solve $V_\alpha(D_i) = \hat{\lambda}_\alpha$ numerically in α to obtain that unit's r-value. Fig. 4 illustrates the computation for two units in the NBA example. Pseudocode for the algorithm and elements of the R package implementation are given in the on-line supplementary material document.

The grey curves in Fig. 4 show, for each of 461 NBA players who attempted at least one free throw in the entire 2013–2014 regular season, the tail probability function $V_\alpha(D_i) = P(\theta_i \geqslant \theta_\alpha | D_i)$; two are highlighted in magenta. Recall that θ_α is such that $P(\theta_i \geqslant \theta_\alpha) = \alpha$; in this case a conjugate beta(a, b) model was fitted to obtain these marginal quantiles ($\hat{a} = 15.12$ and $\hat{b} = 5.38$). At each value of $\alpha(j)$ on a grid, the empirical distribution of $\{V_{\alpha(j)}(D_i)\}$ was computed and reduced to a quantile such that the empirical frequency exceeding the quantile is $\alpha(j)$ (the red dots). We smoothed these to obtain the quantile fuction $\hat{\lambda}_\alpha$ (the blue curve). Two r-values are shown (the broken vertical lines, at r-values 0.016 and 0.488), obtained by solving in α equality of the unit-specific $V_\alpha(D_i)$ and the systemwide $\hat{\lambda}_\alpha$. Scaling by logarithms (horizontal) and square root (vertical) was done to aid visualization.

5. Sampling performance

The r-value is defined using the joint distribution of data D_i and the target parameter θ_i, but it is computed empirically from an estimate of that joint distribution. Accurate distributional estimation may be possible from large-scale data sets, but it is nonetheless useful to investigate how the optimality that is guaranteed by theorems 1 and 2 deteriorates in finite sample situations. Simulations of the normal–normal model show that computed r-values retain their performance benefits compared with other ranking procedures, and thus some uncertainty in the quantile function λ_α or in the distribution of θ_i does not clearly disable the procedure. For example, Fig. 5 shows simulation-based estimates of agreement, $P\{\hat{r}_n(D_i) \leqslant \alpha, \theta_i \geqslant \theta_\alpha\}$, for both the computed r-values $\{\hat{r}_n(D_i)\}$ and for other ranking methods. We adapt the notation to include the sample size n and the *hat* mark to emphasize that the computed r-values involve estimation of the marginal distribution function F of θ_i and the quantile function λ_α. r-value performance is not adversely affected by low sample sizes in this case. Other simulations demonstrate that this superiority is not sensitive to the distribution of variances or to the extent of smoothing that is used to compute quantiles (see the on-line supplementary material, Figs S4 and S5).

A more general consistency property holds for models that are sufficiently regular that the following four conditions are satisfied.

Condition 1. Triples $(\theta_i, X_i, \sigma_i^2)$, for $i = 1, 2, \ldots, n$, are independent and identically distributed from a joint distribution for which θ_i and σ_i^2 are independent and have positive densities f and g with respect to Lebesgue measure on $\mathbb{R}$ and $\mathbb{R}^+$ respectively.

Condition 2. From data $\{D_i = (X_i, \sigma_i^2) : i = 1, 2, \ldots, n\}$, we have an estimator $\hat{F}_n$ of F, where $F(\theta) = \int_{-\infty}^{\theta} f(t)\,\mathrm{d}t$, that is invariant under permutations of the observations. The sequence of distributions converges weakly, $\hat{F}_n \Rightarrow F$, almost surely as $n \to \infty$.

The estimator $\hat{F}_n$ could be parametric or non-parametric (see Lindsay (1995)). For each α, the marginal quantile $\theta_\alpha = F^{-1}(1-\alpha)$ is estimated by $\hat{\theta}_{\alpha,n} = \hat{F}_n^{-1}(1-\alpha)$, and the posterior tail probability, $V_\alpha(x, \sigma^2)$, given a potential data point (x, σ^2), is estimated by

$$\hat{V}_{\alpha,n}(x, \sigma^2) = \int_{\hat{\theta}_{\alpha,n}}^{\infty} p(x|\theta, \sigma^2)\,\mathrm{d}\hat{F}_n(\theta) \Big/ \int_{-\infty}^{\infty} p(x|\theta, \sigma^2)\,\mathrm{d}\hat{F}_n(\theta). \tag{12}$$

Here $p(x|\theta, \sigma^2)$ is the local sampling density, which we consider to have a known form.

Condition 3. The local sampling density satisfies

(a) $p(x|\theta, \sigma^2)$ is continuous in (x, θ, σ^2),
(b) there is a continuous function $K(\sigma^2)$ such that $0 < p(x|\theta, \sigma^2) \leqslant K(\sigma^2)$ for all arguments and,
(c) for any $x_1 > x_0$ and $\sigma^2 > 0$, $p(x_1|\theta, \sigma^2)/p(x_0|\theta, \sigma^2)$ is increasing in θ.

Let $H_\alpha(v) = P\{V_\alpha(X_i, \sigma_i^2) \leqslant v\}$, $\lambda_\alpha = H_\alpha^{-1}(1-\alpha)$ and $t_\alpha^*(\sigma^2) = \inf\{x : V_\alpha(x, \sigma^2) \geqslant \lambda_\alpha\}$.

Condition 4. There are no values of σ^2 and $\alpha_1 \neq \alpha_2$ such that $t_{\alpha_1}^*(\sigma^2) = t_{\alpha_2}^*(\sigma^2)$.

The normal–normal model satisfies condition 1 by design, condition 3 by inspection and condition 4 by theorem 3, and it will satisfy condition 2 for typical parametric or non-parametric estimates of F. Indeed condition 3 is readily verified in many settings, but condition 4 is more difficult because it involves the marginal distribution of local posterior probabilities, which is

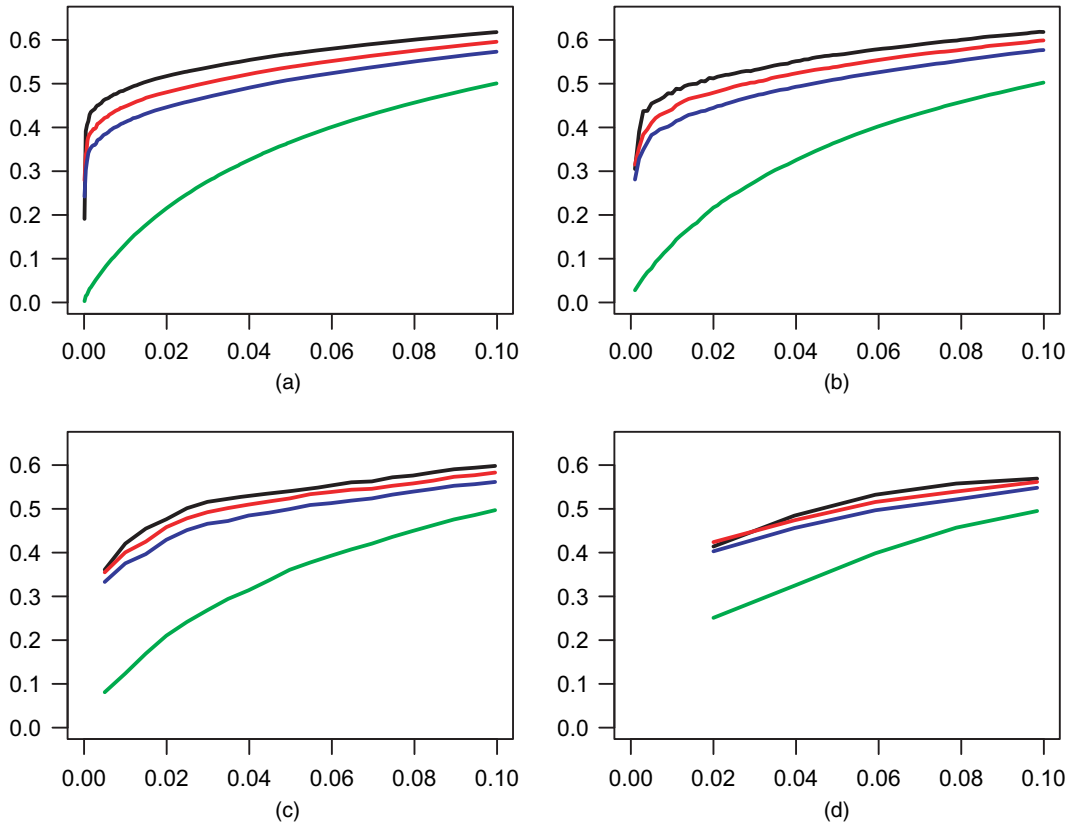

Fig. 5. Finite sample performance of the r-value (——), PM (——), PER (——) and MLE (——) in the normal–normal model (the simulation-based agreement compares the true top α list with the estimated top α list for various methods and for $1/n \leqslant \alpha \leqslant 0.1$ (common horizontal axis), when the marginal distribution of θ_i and the quantile λ_α are both estimated from available data (no smoothing); the common vertical axis is agreement/α; $\sigma_i^2 \sim \text{gamma}(\frac{1}{2}, \frac{1}{2})$, and results from 1000 simulated data sets were averaged for each panel): (a) $n = 10000$; (b) $n = 1000$; (c) $n = 200$; (d) $n = 50$

often analytically intractable. We have confirmed conditions 1–4 in a gamma–inverse gamma model (see the on-line supplementary material).

The ideal r-value $r(D_i) = \inf\{\alpha \in (0,1) : V_\alpha(D_i) \geqslant \lambda_\alpha\}$ is not computable when the underlying distributions are unknown, though model regularity assures that $r(D_i)$ is the unique root (in α) of the equation $V_\alpha(D_i) = \lambda_\alpha$. Approximating $H_\alpha(v)$ we have the empirical distribution function, $\hat{H}_{\alpha,n}(v) = (1/n)\Sigma_{i=1}^n \mathbf{1}\{\hat{V}_{\alpha,n}(X_i, \sigma_i^2) \leqslant v\}$, and the unsmoothed quantile $\hat{\lambda}_{\alpha,n} = \hat{H}_{\alpha,n}^{-1}(1-\alpha) = \inf\{v : \hat{H}_{\alpha,n}(v) \geqslant 1-\alpha\}$. A natural estimate of $r(D_i)$ is $\hat{r}_n(D_i) = \inf\{\alpha \in (0,1) : \hat{V}_{\alpha,n}(D_i) \geqslant \hat{\lambda}_{\alpha,n}\}$. To analyse estimation error, it is helpful to define the related quantity $r^\delta(D_i) = \min[\inf\{\alpha \in [\delta, 1] : V_\alpha(D_i) \geqslant \lambda_\alpha\}, 1-\delta]$ for $\delta \in (0, \frac{1}{2})$, and the sample version, $\hat{r}_n^\delta(D_i) = \min[\inf\{\alpha \in [\delta, 1] : \hat{V}_{\alpha,n}(D_i) \geqslant \hat{\lambda}_{\alpha,n}\}, 1-\delta]$. It happens that $r^\delta(D_i) = r(D_i)$ when both reside in $[\delta, 1-\delta]$; we think of δ as an arbitrarily small value that ameliorates boundary effects in the estimated quantile function $\hat{H}_{\alpha,n}^{-1}$.

Theorem 4. If the model satisfies conditions 1–4 and $n \to \infty$, then for $\delta \in (0, \frac{1}{2})$, and all $\alpha \in [\delta, 1-\delta]$, $(1/n)\Sigma_{i=1}^n \mathbf{1}\{\hat{r}_n^\delta(D_i) \leqslant \alpha\} \to_P \alpha$. Furthermore,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\hat{r}_n(D_i) \leqslant \alpha, \theta_i \geqslant \theta_\alpha\} \geqslant P\{r(D_i) \leqslant \alpha, \theta_i \geqslant \theta_\alpha\} + o_P(1). \tag{13}$$

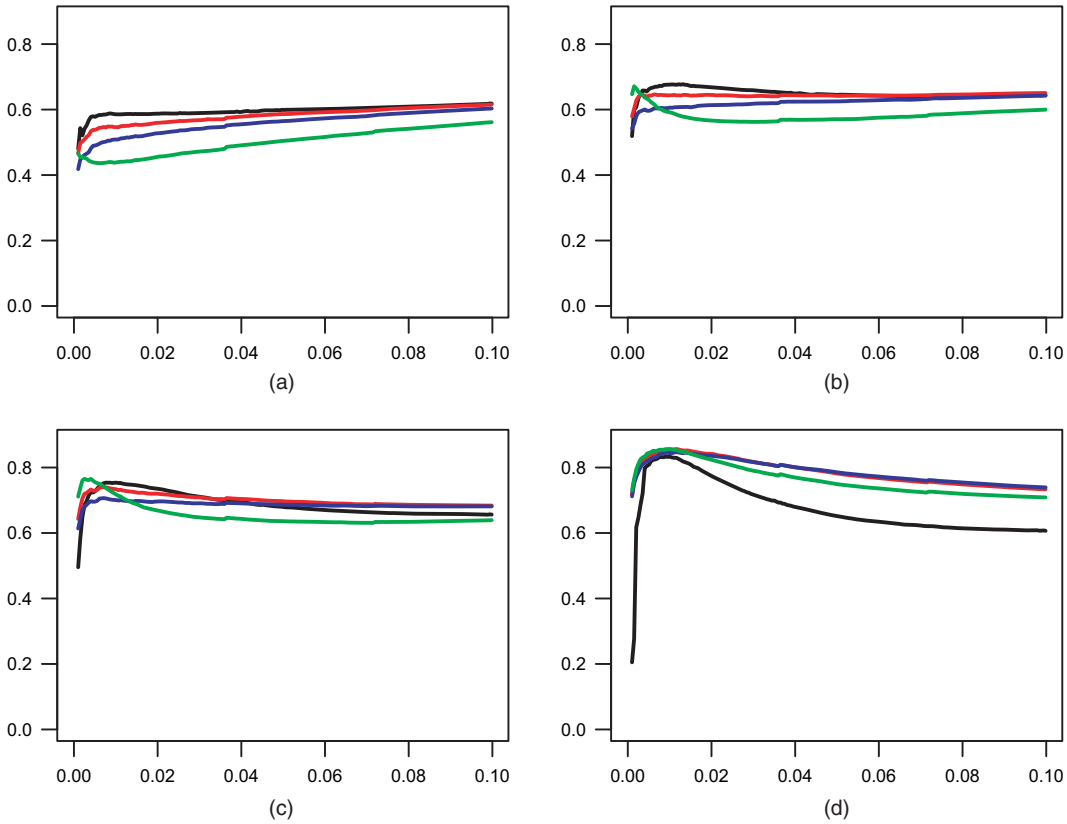

Fig. 6. Effects on agreement of model misspecification for the r-value (——), PM (——), MLE (——) and PER (——) (the r-value performance deteriorates when the true distribution of effects θ_i is much heavier tailed (Student's t on df degrees of freedom) than is used to construct the r-value (normal); the case shown involves $n = 2000$; the axes are as in Fig. 5): (a) df = 6; (b) df = 4; (c) df = 3; (d) df = 2

The quantity $P\{r(D_i) \leqslant \alpha, \theta_i \geqslant \theta_\alpha\}$ is the optimal agreement, as in theorem 2.

Essentially, computed r-values are uniformly distributed and achieve the maximal agreement in large samples as long as the generative distributions are sufficiently regular and consistently estimated.

Model uncertainty can have a bigger effect than system parameter uncertainty on the r-value performance. Fig. 6 shows some reduced performance of r-value in case F is misspecified as normal when it is fact heavier tailed. Other misspecifications may have less effect, such as when the true F is a finite mixture of normal distributions or when there are unmodelled dependences between θ_i and σ_i^2. Examples are provided in the on-line supplementary material, Figs S6–S8. Without pursuing a comparative analysis, we note finally that an alternative estimator of $\lambda_\alpha = H_\alpha^{-1}(1-\alpha)$ may be obtained by working out, perhaps via simulation, the induced distribution of $V_\alpha(D^*)$ for bootstrap data D^* drawn from the fitted model.

6. Discussion

For examples touched on here as well as for many others within the domain of large-scale inference, a basic statistical problem is to rank units and to select the top units by some measure.

Precisely how the output of such inference is to be used depends very much on the context; admittedly we have not focused on these operational issues. For example, the output might trigger follow-up experiments in a genomic study (e.g. Pyeon *et al.* (2007)), it might affect resource allocation in some performance evaluation (e.g. Paddock and Louis (2011)), or it might spark a debate about who really is the best free-throw shooter. Our emphasis on a statistical framework for large-scale ranking and selection responds to evident weaknesses of available methodologies and the potential utility of the proposed *r*-value scheme, especially when there is great variation in the amount of information per unit. Also, where an emphasis of large-scale inference has been on testing and sparsity assumptions, the *r*-value computation addresses a practical problem to organize large numbers of non-null units.

By casting the problem via empirical Bayes methods, we express agreement between true and reported top lists as a certain joint probability that is subject to explicit optimization, taking advantage of an equivalence between ranking and threshold functions (Section 2). Roughly speaking, an *r*-value is a Bayes rule for the binary loss which indicates failure to place the unit correctly in the top α-fraction of units, though to formalize this one requires multiple loss functions and a distributional constraint (Section 3.1). In spite of this connection to Bayesian inference, the *r*-value method seems not to have been previously identified by that reasoning. Theoretical support for the method has been developed here for a measurement model (Sections 2.1–2.3). Establishing that *r*-values maximize agreement in the more general cases that were considered in Section 2.4 remains to be investigated. Where the analysis in Section 2 treats the joint distribution of data and unit level parameters as known, this model must be estimated from systemwide data in each application. We report sufficient conditions for first-order asymptotic correctness (theorem 4). Within-model simulations show good *r*-value performance under a range of conditions (Section 5). Performance deteriorates when the model is misspecified, and we recommend that standard model diagnostics accompany the *r*-value computation. Further investigation is warranted for non-parametric or semiparametric models, as the basic *r*-value statistic does not require a parametric formulation.

Acknowledgements

This research was supported in part by grants from the US National Institutes of Health: R21 HG006568, T32 GM074904 and U54AI117924. The authors thank Christina Kendziorski and several reviewers for critical comments on earlier drafts. Additional details on data analyses, threshold functions and computation are provided in an on-line supplementary material document.

Appendix A: Proofs

A.1. Theorem 1

In this section we assume that all distributions have continuous densities on their support. From the calculus of variations (e.g. Jost and Li-Jost (1998), chapter 1), for a continuously differentiable threshold function $t^* = t^*_\alpha(\sigma^2)$ to maximize agreement (2) subject to the size constraint (1), it must be a critical point of the objective function: $I(t) = \int_0^\infty F(t, \sigma^2)\, g(\sigma^2)\, \mathrm{d}\sigma^2$, where

$$F(t, \sigma^2) = P\{X \geqslant t_\alpha(\sigma^2), \theta \geqslant \theta_\alpha | \sigma^2\} + \lambda P\{X \geqslant t_\alpha(\sigma^2) | \sigma^2\},$$

and where λ is a Lagrange multiplier. Here and to follow we suppress the unit identifier i in the notation for X and σ^2, as we are focusing on a generic unit. The Lagrange–Euler theorem guides us to ignore for a moment that t is a function and to consider derivatives of F in t as a real-valued argument:

$$F_t(t,\sigma^2) := \frac{\mathrm{d}}{\mathrm{d}t} F(t,\sigma^2) = 0 \qquad \text{for all } \sigma^2 \text{ in the support of } g. \tag{14}$$

This Lagrange–Euler equation simplifies:

$$\begin{aligned} F_t(t,\sigma^2) &= \frac{\mathrm{d}}{\mathrm{d}t}\left\{ \int_{\theta_\alpha}^{\infty} P(X \geqslant t|\theta,\sigma^2)\, f(\theta|\sigma^2)\, \mathrm{d}\theta + \lambda\, P(X \geqslant t|\sigma^2) \right\} \\ &= -\, p(t|\sigma^2) \left\{ \int_{\theta_\alpha}^{\infty} \frac{p(t|\theta,\sigma^2)\, f(\theta|\sigma^2)}{p(t|\sigma^2)} \mathrm{d}\theta + \lambda \right\} \\ &= -\, p(t|\sigma^2) \left\{ P(\theta \geqslant \theta_\alpha | X = t, \sigma^2) + \lambda \right\}. \end{aligned}$$

In this development, $p(t|\theta,\sigma^2)$ is the sampling density of X given θ and σ^2 evaluated at the argument t, and similarly $p(t|\sigma^2)$ is the density marginal to θ but conditional on σ^2. Solving $F_t(t,\sigma^2) = 0$ for all $\sigma^2 > 0$ gives result (3).

A.2. Theorem 2

Let α and λ both be fixed in $(0,1)$, and for binary statistics $a = a(X,\sigma^2) \in \{0,1\}$ consider the objective function

$$I_{\alpha,\lambda}(a) = E[a(X,\sigma^2)\{\mathbf{1}(\theta \geqslant \theta_\alpha) - \lambda\}]. \tag{15}$$

Maximizing $I_{\alpha,\lambda}(a)$ is achieved by maximizing the conditional expectation

$$E[a(X,\sigma^2)\{\mathbf{1}(\theta \geqslant \theta_\alpha) - \lambda\}|X,\sigma^2]$$

for every conditioning event, but this conditional expectation is $a(X,\sigma^2)\{V_\alpha(X,\sigma^2) - \lambda\}$, which is maximized at $a^*_{\alpha,\lambda}(X,\sigma^2) = \mathbf{1}\{V_\alpha(X,\sigma^2) \geqslant \lambda\}$. Now we select a particular value λ_α of λ for which $E\{a^*_{\alpha,\lambda}(X,\sigma^2)\} = \alpha$, we denote the resulting rule by $\hat{a}_\alpha = a^*_{\alpha,\lambda_\alpha}$ and we construct the threshold function

$$t^*_\alpha(\sigma^2) = \inf\{x : \hat{a}_\alpha(x,\sigma^2) = 1\} = \inf\{x : V_\alpha(x,\sigma^2) \geqslant \lambda_\alpha\}. \tag{16}$$

By right continuity and monotonicity it follows that $X \geqslant t^*_\alpha(\sigma^2)$ is equivalent to $V_\alpha(X,\sigma^2) \geqslant \lambda_\alpha$. The equivalence will also hold if there are values of σ^2 such that $V_\alpha(x,\sigma^2) < \lambda_\alpha$ for all x or if there are values of σ^2 with $V_\alpha(x,\sigma^2) \geqslant \lambda_\alpha$ for all x, where $t^*_\alpha(\sigma^2)$ is set to ∞ and $-\infty$ respectively. This equivalence implies the size constraint but also allows us to develop a comparison of the thresholds $\{t^*_\alpha\}$ and any other thresholds $\{t_\alpha\}$ which also satisfy that constraint. Using the optimality of $\hat{a}_\alpha$ in equation (15), it follows that

$$I_{\alpha,\lambda_\alpha}(\hat{a}_\alpha) \geqslant I_{\alpha,\lambda_\alpha}(b_\alpha) \tag{17}$$

where $b_\alpha(X,\sigma^2) = \mathbf{1}\{X \geqslant t_\alpha(\sigma^2)\}$ is the threshold-based rule that we are comparing with the putative optimal threshold. Expanding inequality (17),

$$P\{X \geqslant t^*_\alpha(\sigma^2), \theta \geqslant \theta_\alpha\} - \lambda_\alpha P\{X \geqslant t^*_\alpha(\sigma^2)\} \geqslant P\{X \geqslant t_\alpha(\sigma^2), \theta \geqslant \theta_\alpha\} - \lambda_\alpha P\{X \geqslant t_\alpha(\sigma^2)\}$$

from which optimality of $\{t^*_\alpha\}$ follows immediately, since both marginal probabilities involved equal α.

A.3. Theorem 3

Suppose that there is crossing, in contradiction to the claim, i.e. there exists $(\alpha_1, \alpha_2, \sigma_0^2)$ with $\alpha_1 < \alpha_2$ such that $t_{\alpha_1}(\sigma_0^2) = t_{\alpha_2}(\sigma_0^2)$. By the mean value theorem, there exists $c \in [\alpha_1, \alpha_2]$ such that $\partial t_\alpha(\sigma_0^2)/\partial\alpha|_{\alpha=c} = \{t_{\alpha_2}(\sigma_0^2) - t_{\alpha_1}(\sigma_0^2)\}/(\alpha_2 - \alpha_1) = 0$, which is in violation of the derivative condition.

In the normal–normal model, $t^*_\alpha(\sigma^2) = \theta_\alpha(\sigma^2+1) - u_\alpha\sqrt{\{\sigma^2(\sigma^2+1)\}}$ as presented in equation (4), with $\theta_\alpha = \Phi^{-1}(1-\alpha)$, u_α defined by the constraint equation (5) and Φ the cumulative distribution function of the standard normal distribution. Our proof that this threshold has a negative derivative in α uses the interesting fact that $h(a) = \phi\{\Phi^{-1}(a)\}$ is strictly concave for $a \in (0,1)$, which may be confirmed by differentiation. (Here ϕ is the density function that is associated with Φ.)

Lemma 1. In the normal–normal model, assuming that $P(\sigma^2 = 0) < 1$, we have $\mathrm{d}u_\alpha/\mathrm{d}\alpha > \mathrm{d}\theta_\alpha/\mathrm{d}\alpha$.

Proof. Let

$$D_\alpha(\sigma^2) = \Phi\{\theta_\alpha\sqrt{(\sigma^2+1)} - u_\alpha\sigma\}, \tag{18}$$

so that $E\{D_\alpha(\sigma^2)\} = 1 - \alpha$ is the constraint equation (5). Suppose, by contradiction, that $-u'_\alpha \geqslant -\theta'_\alpha$,

where primes indicate differentiation with respect to α. Differentiating equation (5) with respect to α, and using G to denote the distribution function of σ_i^2, we obtain

$$\begin{aligned}
1 &= -\theta'_\alpha \int_0^\infty \sqrt{(\sigma^2+1)}\,\phi[\Phi^{-1}\{D_\alpha(\sigma^2)\}]\,\mathrm{d}G(\sigma^2) - (-u'_\alpha)\int_0^\infty \sigma\,\phi[\Phi^{-1}\{D_\alpha(\sigma^2)\}]\,\mathrm{d}G(\sigma^2) \\
&\leqslant -\theta'_\alpha \int_0^\infty \sqrt{(\sigma^2+1)}\,\phi[\Phi^{-1}\{D_\alpha(\sigma^2)\}]\,\mathrm{d}G(\sigma^2) + \theta'_\alpha \int_0^\infty \sigma\,\phi[\Phi^{-1}\{D_\alpha(\sigma^2)\}]\,\mathrm{d}G(\sigma^2) \\
&= -\theta'_\alpha \int_0^\infty \{\sqrt{(\sigma^2+1)}-\sigma\}\,\phi[\Phi^{-1}\{D_\alpha(\sigma^2)\}]\,\mathrm{d}G(\sigma^2) \\
&< -\theta'_\alpha \int_0^\infty \phi[\Phi^{-1}\{D_\alpha(\sigma^2)\}]\,\mathrm{d}G(\sigma^2) \qquad \text{unless } P(\sigma^2=0)=1 \\
&= -\theta'_\alpha E[h\{D_\alpha(\sigma^2)\}]. \qquad (19)
\end{aligned}$$

From Jensen's inequality, we know that $E[h\{D_\alpha(\sigma^2)\}] \leqslant h[E\{D_\alpha(\sigma^2)\}]$. Hence,

$$1 < -\theta'_\alpha h[E\{D_\alpha(\sigma^2)\}] = -\theta'_\alpha h(1-\alpha) = -\theta'_\alpha \phi\{\Phi^{-1}(1-\alpha)\} = 1.$$

This contradiction leads us to conclude that $-u'_\alpha < -\theta'_\alpha$, thus establishing lemma 1.

To complete the non-crossing proof, we differentiate equation (4) in α:

$$\begin{aligned}
\frac{\partial t^*_\alpha(\sigma^2)}{\partial\alpha} &= (\sigma^2+1)\left\{\frac{\mathrm{d}\theta_\alpha}{\mathrm{d}\alpha} - \frac{\mathrm{d}u_\alpha}{\mathrm{d}\alpha}\sqrt{\left(\frac{\sigma^2}{\sigma^2+1}\right)}\right\} \\
&< (\sigma^2+1)\left\{\frac{\mathrm{d}\theta_\alpha}{\mathrm{d}\alpha} - \frac{\mathrm{d}\theta_\alpha}{\mathrm{d}\alpha}\sqrt{\left(\frac{\sigma^2}{\sigma^2+1}\right)}\right\} \\
&= \frac{\mathrm{d}\theta_\alpha}{\mathrm{d}\alpha}(\sigma^2+1)\left\{1-\sqrt{\left(\frac{\sigma^2}{\sigma^2+1}\right)}\right\} < 0,
\end{aligned}$$

where the first inequality comes from lemma 1 and the second from the fact that $\mathrm{d}\theta_\alpha/\mathrm{d}\alpha = -1/\phi(\theta_\alpha) < 0$. For the trivial case when $P(\sigma^2=0)=1$, we note that the optimal 'threshold function' is $t^*_\alpha(0)=\theta_\alpha$ which obviously satisfies $\partial t^*_\alpha(\sigma^2)/\partial\alpha < 0$.

A.4. Theorem 4

We proceed in steps.

Lemma 2. Assume that conditions 1 and 2 hold. For each $\alpha \in (0,1)$, $\hat{\theta}_{\alpha,n} \to \theta_\alpha$ almost surely as $n \to \infty$.

Proof. At continuity points p of F^{-1}, $\hat{F}_n^{-1}(p)$ converges almost surely to the limiting quantile $F^{-1}(p)$ by condition 2 and, for example, lemma 21.2 of van der Vaart (1998). Continuity of F^{-1} follows from condition 1, and thus the result follows.

Lemma 3. Assume conditions 1–3. The limiting posterior tail probability $V_\alpha(X_i,\sigma_i^2)$ is continuous and non-decreasing in α for any data (X_i,σ_i^2). Further, as $n\to\infty$,

$$\sup_{\alpha\in(0,1)} |\hat{V}_{\alpha,n}(X_i,\sigma_i^2) - V_\alpha(X_i,\sigma_i^2)| \to 0 \qquad \text{almost surely.}$$

Proof. First we confirm pointwise (in α) convergence of the numerator and the denominator of equation (12) when evaluated at $(X_i,\sigma_i^2)=(x,\sigma^2)$. The denominator is immediate, owing to $p(x|\theta,\sigma^2)$ being bounded and continuous in θ, and owing to the almost sure weak convergence of $\hat{F}_n$. For the numerator, note that the mapping $\theta \mapsto \mathbf{1}(\theta \geqslant \theta_\alpha)\,p(x|\theta,\sigma^2)$, for fixed (x,σ^2), is continuous except at θ_α, which has zero point mass in the limiting distribution F. Thus $\int_{\theta_\alpha}^\infty p(x|\theta,\sigma^2)\,\mathrm{d}\hat{F}_n(\theta)$ converges almost surely to $\int_{\theta_\alpha}^\infty p(x|\theta,\sigma^2)\,\mathrm{d}F(\theta)$, using condition 2 and, for example, theorem 2.3 of van der Vaart (1998). It is sufficient to confirm that the error e_n, defined as $e_n = |\int_{\hat{\theta}_{\alpha,n}}^\infty p(x|\theta,\sigma^2)\,\mathrm{d}\hat{F}_n(\theta) - \int_{\theta_\alpha}^\infty p(x|\theta,\sigma^2)\,\mathrm{d}\hat{F}_n(\theta)|$, converges almost surely to 0. With the bound condition 3, part (b), and taking any $\epsilon > 0$, we have

$$e_n \leqslant K(\sigma^2)\int_{\mathbb{R}} |\mathbf{1}(\theta \geqslant \hat{\theta}_{\alpha,n}) - \mathbf{1}(\theta \geqslant \theta_\alpha)|\,\mathrm{d}\hat{F}_n(\theta) \leqslant K(\sigma^2)\int_{\theta_\alpha-\epsilon}^{\theta_\alpha+\epsilon} \mathrm{d}\hat{F}_n(\theta) \qquad \text{for } n \geqslant N_\epsilon,$$

where the second inequality is almost sure owing to lemma 2. Consequently, $\limsup_n e_n$ is almost surely bounded by $K(\sigma^2)\{F(\theta_\alpha+\epsilon)-F(\theta_\alpha-\epsilon)\}$ for every $\epsilon>0$, and so, pointwise in α, the limiting error must be 0, since F contains no atoms. On continuity and monotonicity of $\alpha\mapsto V_\alpha(D_i)$, let $\{\alpha_n\}$ denote a sequence in $(0,1)$ for which $\alpha_n\geqslant\alpha$. We have

$$\begin{aligned}0\leqslant V_{\alpha_n}(D_i)-V_\alpha(D_i)&=\frac{1}{p(D_i)}\int_{\theta_{\alpha_n}}^{\theta_\alpha}p(X_i|\theta_i,\sigma_i^2)\,\mathrm{d}F(\theta_i)\\&\leqslant\frac{1}{p(D_i)}K(\sigma_i^2)(\alpha_n-\alpha).\end{aligned}$$

Monotonicity is immediate from this, but also, if $\alpha_n\to\alpha$, we obtain right continuity of $V_{\alpha_n}(D_i)$. A comparable argument gives left continuity. Uniform convergence follows from Polya's theorem (e.g. Bickel and Millar (1992)).

Lemma 4. If condition 3 holds, the mappings $(x,\sigma^2)\mapsto V_\alpha(x,\sigma^2)$ and $(x,\sigma^2)\mapsto\hat{V}_{\alpha,n}(x,\sigma^2)$ are continuous. Further, for any $x_1>x_0$, $V_\alpha(x_1,\sigma^2)>V_\alpha(x_0,\sigma^2)$ and $\hat{V}_{\alpha,n}(x_1,\sigma^2)>\hat{V}_{\alpha,n}(x_0,\sigma^2)$.

Proof. Take a sequence $\{d_m=(x_m,\sigma_m^2)\}$ with $d_m\to d=(x,\sigma^2)$, and observe that, for each n,

$$\lim_{m\to\infty}\int_{-\infty}^{\infty}p(x_m|\theta,\sigma_m^2)\,\mathrm{d}\hat{F}_n(\theta)=\int_{-\infty}^{\infty}\lim_{m\to\infty}p(x_m|\theta,\sigma_m^2)\,\mathrm{d}\hat{F}_n(\theta)=\int_{-\infty}^{\infty}p(x|\theta,\sigma^2)\,\mathrm{d}\hat{F}_n(\theta).$$

The first equality follows from a dominated convergence argument, using condition 3, part (b), and the second equality follows from continuity of the local sampling density, condition 3, part (a). The same would hold if we replaced the integrand $p(x_m|\theta,\sigma_m^2)$ with $\mathbf{1}(\theta\geqslant\hat{\theta}_{\alpha,n})\,p(x_m|\theta,\sigma_m^2)$ and likewise modified the limit. Thus continuity of the ratio $\hat{V}_{\alpha,n}(x,\sigma^2)$ is established. The argument for $V_\alpha(x,\sigma^2)$ is analogous.

On the monotonicity claim, note that $V_\alpha(x,\sigma^2)=1/\{1+1/\psi(x)\}$, where

$$\psi(x)=\int_{\theta_\alpha}^{\infty}p(x|\theta,\sigma^2)\,\mathrm{d}F(\theta)\Big/\int_{-\infty}^{\theta_\alpha}p(x|\theta,\sigma^2)\,\mathrm{d}F(\theta).$$

Showing that $\psi(x)$ is increasing would be enough to prove that $V_\alpha(x,\sigma^2)$ is increasing. Write

$$\frac{\psi(x_1)}{\psi(x_0)}=\frac{\displaystyle\int_{\theta_\alpha}^{\infty}p(x_1|\theta,\sigma^2)\,\mathrm{d}F(\theta)}{\displaystyle\int_{\theta_\alpha}^{\infty}p(x_0|\theta,\sigma^2)\,\mathrm{d}F(\theta)}\frac{\displaystyle\int_{-\infty}^{\theta_\alpha}p(x_0|\theta,\sigma^2)\,\mathrm{d}F(\theta)}{\displaystyle\int_{-\infty}^{\theta_\alpha}p(x_1|\theta,\sigma^2)\,\mathrm{d}F(\theta)}=y_1y_2.\tag{20}$$

If we let $\rho(\theta)=p(x_1|\theta,\sigma^2)/p(x_0|\theta,\sigma^2)$, then, because $\rho(\theta)$ is increasing by condition 3,

$$\begin{aligned}y_1=\frac{\displaystyle\int_{\theta_\alpha}^{\infty}p(x_1|\theta,\sigma^2)\,\mathrm{d}F(\theta)}{\displaystyle\int_{\theta_\alpha}^{\infty}p(x_0|\theta,\sigma^2)\,\mathrm{d}F(\theta)}&=\frac{\displaystyle\int_{\theta_\alpha}^{\infty}p(x_0|\theta,\sigma^2)\rho(\theta)\,\mathrm{d}F(\theta)}{\displaystyle\int_{\theta_\alpha}^{\infty}p(x_0|\theta,\sigma^2)\,\mathrm{d}F(\theta)}\\&>\frac{\rho(\theta_\alpha)\displaystyle\int_{\theta_\alpha}^{\infty}p(x_0|\theta,\sigma^2)\,\mathrm{d}F(\theta)}{\displaystyle\int_{\theta_\alpha}^{\infty}p(x_0|\theta,\sigma^2)\,\mathrm{d}F(\theta)}=\rho(\theta_\alpha).\end{aligned}$$

Likewise,

$$\begin{aligned}y_2=\frac{\displaystyle\int_{-\infty}^{\theta_\alpha}p(x_0|\theta,\sigma^2)\,\mathrm{d}F(\theta)}{\displaystyle\int_{-\infty}^{\theta_\alpha}p(x_1|\theta,\sigma^2)\,\mathrm{d}F(\theta)}&=\frac{\displaystyle\int_{-\infty}^{\theta_\alpha}p(x_0|\theta,\sigma^2)\,\mathrm{d}F(\theta)}{\displaystyle\int_{-\infty}^{\theta_\alpha}p(x_0|\theta,\sigma^2)\rho(\theta)\,\mathrm{d}F(\theta)}\\&>\frac{\displaystyle\int_{-\infty}^{\theta_\alpha}p(x_0|\theta,\sigma^2)\,\mathrm{d}F(\theta)}{\rho(\theta_\alpha)\displaystyle\int_{-\infty}^{\theta_\alpha}p(x_0|\theta,\sigma^2)\,\mathrm{d}F(\theta)}=\frac{1}{\rho(\theta_\alpha)}.\end{aligned}$$

Hence, $y_1 y_2 > 1$ and from equation (20) we know that $\psi(x_1)/\psi(x_0) > 1$, whence $\psi(x)$ is increasing. The argument for monotonicity of $\hat{V}_{\alpha,n}(x, \sigma^2)$ is completely analogous.

Lemma 5. Let $B = (0, 1) \times (0, 1)$ denote the open unit square. Assume conditions 1–3, and let $\hat{H}^{-1}_{\alpha,n}(p)$ be the quantile function associated with the empirical distribution of $\{\hat{V}_{\alpha,n}(X_i, \sigma_i^2)\}$, for some $(\alpha, p) \in B$. As $n \to \infty$, $\hat{H}^{-1}_{\alpha,n}(p) \to_{\mathrm{P}} H^{-1}_{\alpha}(p)$, and this limit is continuous on B. The limit function and each estimate are non-decreasing in p for each α and non-decreasing in α for each p. Furthermore, the convergence is uniform on any closed square $A_\delta = [\delta, 1-\delta] \times [\delta, 1-\delta]$ for $\delta \in (0, \frac{1}{2})$.

Proof. To simplify the notation, let $\xi_i = V_\alpha(X_i, \sigma_i^2)$ and $\hat{\xi}_{n,i} = \hat{V}_{\alpha,n}(X_i, \sigma_i^2)$. For $v \in (0, 1)$, we consider the intermediate empirical distribution $\tilde{H}_{\alpha,n}(v) = (1/n)\sum_{i=1}^n \mathbf{1}(\xi_i \leqslant v)$, which entails no estimation error in $\hat{F}_n$ compared with the computable $\hat{H}_{\alpha,n}(v) = (1/n)\sum_{i=1}^n \mathbf{1}(\hat{\xi}_{n,i} \leqslant v)$, and which converges to $H_\alpha(v)$ by the law of large numbers, owing to condition 1. To show that $\hat{H}_{\alpha,n}$ converges, further define $\Delta_n = |\hat{H}_{\alpha,n}(v) - \tilde{H}_{\alpha,n}(v)|$. For any $\epsilon > 0$ we have

$$\begin{aligned}\Delta_n &\leqslant \frac{1}{n}\sum_{i=1}^n |\mathbf{1}(\hat{\xi}_{n,i} \leqslant v) - \mathbf{1}(\xi_i \leqslant v)| \\ &= \frac{1}{n}\sum_{i=1}^n |\mathbf{1}(\hat{\xi}_{n,i} \leqslant v) - \mathbf{1}(\xi_i \leqslant v)|(1 - U_{n,i}) + \frac{1}{n}\sum_{i=1}^n |\mathbf{1}(\hat{\xi}_{n,i} \leqslant v) - \mathbf{1}(\xi_i \leqslant v)|U_{n,i}\end{aligned}$$

where $U_{n,i} = \mathbf{1}(|\xi_i - \hat{\xi}_{n,i}| > \epsilon)$. Thus

$$\begin{aligned}\Delta_n &\leqslant \frac{1}{n}\sum_{i=1}^n |\mathbf{1}(\hat{\xi}_{n,i} \leqslant v) - \mathbf{1}(\xi_i \leqslant v)|(1 - U_{n,i}) + \frac{1}{n}\sum_{i=1}^n U_{n,i} \\ &\leqslant \frac{1}{n}\sum_{i=1}^n \mathbf{1}(\hat{\xi}_{n,i} \in (v, v+\epsilon]) + \frac{1}{n}\sum_{i=1}^n \mathbf{1}(\xi_i \in (v, v+\epsilon]) + \frac{1}{n}\sum_{i=1}^n U_{n,i}.\end{aligned}$$

From the symmetry of $\hat{F}_n$ in condition 2, all $\hat{\xi}_{n,i}$ are identically distributed, and hence, taking expectations,

$$E(\Delta_n) \leqslant P\{\xi_{n,1} \in (v, v+\epsilon]\} + P\{\xi_1 \in (v, v+\epsilon]\} + E(U_{n,1}). \tag{21}$$

As $n \to \infty$, the term $E(U_{n,1})$ converges to 0 by lemma 3, and likewise the upper bound in expression (21) converges to $2\,P\{\xi_1 \in (v, v+\epsilon]\} = 2\{H_\alpha(v+\epsilon) - H_\alpha(v)\}$. Because $\epsilon > 0$ could be arbitrarily small, and using the continuity of H_α (see lemma S1 in the on-line supplementary material), it follows that $\Delta_n \to_{\mathrm{P}} 0$ and hence $\hat{H}_{\alpha,n}(v) \to_{\mathrm{P}} H_\alpha(v)$ as $n \to \infty$. Convergence in probability of $\hat{H}^{-1}_{\alpha,n}(p)$ to $H^{-1}_{\alpha}(p)$ follows from a basic fact about distributions (see lemma S2 in the supplementary material). Continuity of the limit $H^{-1}_{\alpha}(p)$ on B and co-ordinatewise monotonicity follow from the model regularity conditions (see lemma S2 in the supplementary material). Interestingly, there are two discontinuities on the closed square, at $(0, 1)$ and $(1, 0)$, where the function switches immediately between 0 and 1. Thus we avoid having α near the boundary in establishing uniformity of convergence, which itself follows from a two-dimensional version of Polya's theorem, owing to co-ordinatewise monotonicity and continuity of the limit (see lemma S3 in the supplementary material).

Lemma 6. Define $g_\alpha(D_i) = V_\alpha(D_i) - \lambda_\alpha$ and $\hat{g}_{\alpha,n}(D_i) = \hat{V}_{\alpha,n}(D_i) - \hat{\lambda}_{\alpha,n}$, and assume conditions 1–4. Both $\sup_{\alpha \in [\delta, 1-\delta]} |\hat{\lambda}_{\alpha,n} - \lambda_\alpha|$ and $\sup_{\alpha \in [\delta, 1-\delta]} |\hat{g}_{\alpha,n}(D_i) - g_\alpha(D_i)|$ converge to 0 in probability as $n \to \infty$, for any fixed $\delta \in (0, \frac{1}{2})$. Further, $\alpha \mapsto g_\alpha(D_i)$ is continuous and $g_\alpha(D_i) = 0$ has a unique root $r(D_i)$.

Proof. Let $A_\delta = [\delta, 1-\delta] \times [\delta, 1-\delta]$ denote a closed square within B, and note that

$$\sup_{\alpha \in [\delta, 1-\delta]} |\hat{\lambda}_{\alpha,n} - \lambda_\alpha| = \sup_{\alpha \in [\delta, 1-\delta]} |\hat{H}^{-1}_{\alpha,n}(1-\alpha) - H^{-1}_{\alpha}(1-\alpha)| \leqslant \sup_{(\alpha,p) \in A_\delta} |\hat{H}^{-1}_{\alpha,n}(p) - H^{-1}_{\alpha}(p)|.$$

Uniform convergence of $\hat{\lambda}_{\alpha,n}$ follows from lemma 5. Similarly, uniform convergence of $\hat{g}_{\alpha,n}(D_i)$ follows after also invoking lemma 3.

Continuity of $g_\alpha(D_i)$ follows from lemmas 3 and 5. We deduce uniqueness in the α-root of $g_\alpha(D_i) = 0$ first by noting that condition 1 and continuity of V_α in data (lemma 4) imply the existence of λ_α satisfying condition (6). Were there not at least one α-value for which $g_\alpha(D_i) = 0$, then either $V_\alpha(X_i, \sigma_i^2)$ would always exceed λ_α or it would always be dominated by it. Take the second case; the first is analogous. Find an open ball around $D_i = (X_i, \sigma_i^2)$ such that $g_\alpha(d) < 0$ for all d in this ball and for all α. This ball has some positive

probability, say $\epsilon > 0$, and so $P\{V_\alpha(D_i) < \lambda_\alpha\} \geqslant \epsilon$. But, owing to condition (6), we have a contradiction when $\alpha > 1 - \epsilon$, implying that there must be at least one root of $g_\alpha(D_i) = 0$. The conditions of theorem 2 are met, and so, from continuity in lemma 4, $V_\alpha\{t^*_\alpha(\sigma_i^2), \sigma_i^2\} = \lambda_\alpha$ defines the optimal threshold. Finally, by condition 4, $X_i = t^*_\alpha(\sigma_i^2)$ at exactly one value of α.

Lemma 7. If conditions 1–4 hold, then, for $\delta \in (0, \frac{1}{2})$, $\hat{r}_n^\delta(D_i) \to_{\mathrm{P}} r^\delta(D_i)$ as $n \to \infty$.

Proof. The empirical r-value $\hat{r}_n^\delta(D_i)$ is like a root of $\hat{g}_{\alpha,n}(D_i) = 0$, at least truncated away from end points 0 and 1 but, owing to the sample quantile estimation, $\hat{g}_{\alpha,n}(D_i)$ is not continuous at all α and may admit multiple roots. In spite of this, lemma 6 assures not only continuity of the limit $g_\alpha(D_i)$ having a unique root $r(D_i)$, but also uniform convergence of sample functions $\hat{g}_{\alpha,n}(D_i)$ to this limit, at least on compact subsets of $(0, 1)$. From the first of these properties, the extreme value theorem implies that, for any sufficiently small $\epsilon > 0$, there exists $\nu = \nu(D_i) > 0$ such that the limit function $g_\alpha(D_i)$ has magnitude at least ν for all α with $|r(D_i) - \alpha| > \epsilon$. From the uniform convergence, $|\hat{g}_{\alpha,n}(D_i) - g_\alpha(D_i)| < \nu/2$ with high probability for large n, uniformly for $\alpha \in [\delta, 1 - \delta]$, and thus in this event $|\hat{r}_n^\delta(D_i) - r(D_i)| < \epsilon$. Lemma S4 in the on-line supplementary material provides further details.

Proceeding to prove theorem 4, we know from the unique root result in lemma 6 that events $[r(D_i) \leqslant \alpha]$ and $[V_\alpha(D_i) \geqslant \lambda_\alpha]$ are equivalent, and so the ideal $r(D_i)$ has a uniform(0, 1) distribution by condition (6). The first claim follows from lemma 7 and, for example, theorem 2.3 from van der Vaart (1998), using the fact that $\mathbf{1}\{r^\delta(D_i) \leqslant \alpha\} = \mathbf{1}\{r(D_i) \leqslant \alpha\}$ for $\alpha \in [\delta, 1 - \delta]$. r-values in inequality (13) do not involve truncation away from end points 0 or 1. The claimed lower bound $A_\alpha := P\{r(D_i) \leqslant \alpha, \theta_i \geqslant \theta_\alpha\}$ is maximal because the conditions of theorem 2 are satisfied (lemma 4), and because the maximal agreement is achieved by using $r(D_i)$ (from the unique root remarks above). To establish the bound, let $\hat{A}_{\alpha,n}$ denote the left-hand side of equation (13), and introduce $\tilde{A}_{\alpha,n} = (1/n)\Sigma_{i=1}^n \mathbf{1}[r(D_i) \leqslant \alpha, \theta_i \geqslant \theta_\alpha\}$. Of course $\tilde{A}_{\alpha,n} \to_{\mathrm{P}} A_\alpha$ by the law of large numbers, so at issue are deviations between $\hat{A}_{\alpha,n}$ and $A_{\alpha,n}$ caused by estimation errors. With $\alpha \in [\delta, 1 - \delta]$, $\hat{r}_n^\delta(D_i) \leqslant \alpha$ implies that $\hat{r}_n(D_i) \leqslant \alpha$, and therefore $\hat{A}_{\alpha,n} \geqslant (1/n)\Sigma_{i=1}^n \mathbf{1}\{\hat{r}_n^\delta(D_i) \leqslant \alpha, \theta_i \geqslant \theta_\alpha\}$. Now decompose this lower bound into $\tilde{A}_{\alpha,n} + e_n$, where $e_n = (1/n)\Sigma_{i=1}^n [\mathbf{1}\{\hat{r}_n^\delta(D_i) \leqslant \alpha\} - \mathbf{1}\{r(D_i) \leqslant \alpha\}]\mathbf{1}(\theta_i \geqslant \theta_\alpha)$, using the fact that $\mathbf{1}\{r^\delta(D_i) \leqslant \alpha\} = \mathbf{1}\{r(D_i) \leqslant \alpha\}$ for $\alpha \in [\delta, 1 - \delta]$. Having convergence of e_n in probability to 0 would complete the proof. We have

$$|e_n| \leqslant \frac{1}{n} \sum_{i=1}^n \mathbf{1}(\theta_i \geqslant \theta_\alpha) |\mathbf{1}\{\hat{r}_n^\delta(D_i) \leqslant \alpha\} - \mathbf{1}\{r^\delta(D_i) \leqslant \alpha\}|.$$

By the identical distribution of terms, induced by permutation invariance (condition 2),

$$\begin{aligned} E|e_n| &\leqslant E[\mathbf{1}(\theta_i \geqslant \theta_\alpha) |\mathbf{1}\{\hat{r}_n^\delta(D_1) \leqslant \alpha\} - \mathbf{1}\{r^\delta(D_1) \leqslant \alpha\}|] \\ &\leqslant \sqrt{\alpha}\sqrt{[E|\mathbf{1}\{\hat{r}_n^\delta(D_1) \leqslant \alpha\} - \mathbf{1}\{r^\delta(D_1) \leqslant \alpha\}|]}, \end{aligned}$$

with the second inequality by the Cauchy–Schwartz inequality. The integrand within the expectation on the right-hand side is bounded by 1 and converges in probability to 0 (lemma 7 and theorem 2.3, van der Vaart (1998)), and so $e_n \to_{\mathrm{P}} 0$, completing the proof.

References

Berger, J. O. and Deely, J. (1988) A Bayesian approach to ranking and selection of related means with alternatives to analysis-of-variance methodology. *J. Am. Statist. Ass.*, **83**, 364–373.

Bickel, P. J. and Millar, P. W. (1992) Uniform convergence of probability measures on classes of functions. *Statist. Sin.*, **2**, 1–15.

Brijs, T., Karlis, D., Van den Bossche, F. and Wets, G. (2007) A Bayesian model for ranking hazardous road sites. *J. R. Statist. Soc.* A, **170**, 1001–1017.

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D. and Calus, M. P. (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, **193**, 327–345.

Coelho, C. A. and Mexia, J. T. (2007) On the distribution of the product and ratio of independent generalized gamma-ratio random variables. *Sankhya* A, **69**, 221–255.

Efron, B. (2010) *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, vol. 1. Cambridge: Cambridge University Press.

Entertainment and Sports Programming Network (2014) NBA player free-throw shooting statistics—2013-2014. Entertainment and Sports Programming Network. (Available from `http://espn.go.com/nba/statistics/player/_/stat/free-throws/`.)

Gelman, A. and Price, P. N. (1999) All maps of parameter estimates are misleading. *Statist. Med.*, **18**, 3221–3234.
Gibbons, J. D., Olkin, I. and Sobel, M. (1979) An introduction to ranking and selection. *Am. Statistn*, **33**, 185–195.
Hall, P. and Miller, H. (2010) Modeling the variability of rankings. *Ann. Statist.*, **38**, 2652–2677.
Hao, L., He, Q., Wang, Z., Craven, M., Newton, M. A. and Ahlquist, P. (2013) Limited agreement of independent RNAi screens for virus-required host genes owes more to false-negative than false-positive factors. *PLOS Computnl Biol.*, **9**, article e1003235.
Jost, J. and Li-Jost, X. (1998) *Calculus of Variations*. Cambridge: Cambridge University Press.
Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *J. Am. Statist. Ass.*, **90**, 773–795.
Kendziorski, C., Newton, M., Lan, H. and Gould, M. (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statist. Med.*, **22**, 3899–3914.
Laird, N. M. and Louis, T. A. (1989) Empirical Bayes ranking methods. *J. Educ. Behav. Statist.*, **14**, 29–46.
Lehmann, E. (1986) *Testing Statistical Hypotheses*, 2nd edn. New York: Wiley.
Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M. and Kendziorski, C. (2013) EBSeq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, **29**, 1035–1043.
Lin, R., Louis, T. A., Paddock, S. M. and Ridgeway, G. (2006) Loss function based ranking in two-stage, hierarchical models. *Baysn Anal.*, **1**, 915–946.
Lindsay, B. G. (1995) *Mixture Models: Theory, Geometry and Applications*. Hayward: Institute of Mathematical Statistics.
McCarthy, D. J. and Smyth, G. K. (2009) Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, **25**, 765–771.
Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., Mahajan, A., Prokopenko, I., Kang, H. M., Dina, C., Esko, T., Fraser, R. M., Kanoni, S., Kumar, A., Lagou, V., Langenberg, C., Luan, J., Lindgren, C. M., Müller-Nurasyid, M., Pechlivanis, S., Rayner, N. W., Scott. L. J., Wiltshire, S., Yengo, L., Kinnunen, L., Rossin, E. J., Raychaudhuri, S., Johnson, A. D., Dimas, A. S., Loos, R. J., Vedantam, S., Chen, H., Florez, J. C., Fox, C., Liu, C. T., Rybin, D., Couper, D. J., Kao, W. H., Li, M., Cornelis, M. C., Kraft, P., Sun, Q., van Dam, R. M., Stringham, H. M., Chines, P. S., Fischer, K., Fontanillas, P., Holmen, O. L., Hunt, S. E., Jackson, A. U., Kong, A., Lawrence, R., Meyer, J., Perry, J. R., Platou, C. G., Potter, S., Rehnberg, E., Robertson, N., Sivapalaratnam, S., Stancàkovà, A., Stirrups, K., Thorleifsson, G., Tikkanen, E., Wood, A. R., Almgren, P., Atalay, M., Benediktsson, R., Bonnycastle, L. L., Burtt, N., Carey, J., Charpentier, G., Crenshaw, A. T., Doney, A. S., Dorkhan, M., Edkins, S., Emilsson, V., Eury, E., Forsen, T., Gertow, K., Gigante, B., Grant, G. B., Groves, C. J., Guiducci, C., Herder, C., Hreidarsson, A. B., Hui, J., James, A., Jonsson, A., Rathmann, W., Klopp, N., Kravic, J., Krjutskov, K., Langford, C., Leander, K., Lindholm, E., Lobbens, S., Männistö, S., Mirza, G., Mühleisen, T. W., Musk, B., Parkin, M., Rallidis, L., Saramies, J., Sennblad, B., Shah, S., Sigurðsson, G., Silveira, A., Steinbach, G., Thorand, B., Trakalo, J., Veglia, F., Wennauer, R., Winckler, W., Zabaneh, D., Campbell, H., van Duijn, C., Uitterlinden, A. G., Hofman, A., Sijbrands, E., Abecasis, G. R., Owen, K. R., Zeggini, E., Trip, M. D., Forouhi, N. G., Syvänen, A. C., Eriksson, J. G., Peltonen, L., Nöthen, M. M., Balkau, B., Palmer, C. N., Lyssenko, V.,Tuomi, T., Isomaa, B., Hunter, D. J., Qi, L., Wellcome Trust Case Control Consortium, Meta-Analyses of Glucose and Insulin-related Traits Consortium (MAGIC) Investigators, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Asian Genetic Epidemiology Network-Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Shuldiner, A. R., Roden, M., Barroso, I., Wilsgaard, T., Beilby, J., Hovingh, K., Price, J. F., Wilson, J. F., Rauramaa, R., Lakka, T. A., Lind, L., Dedoussis, G., Njølstad, I., Pedersen, N. L., Khaw, K. T., Wareham, N. J., Keinanen-Kiukaanniemi, S. M., Saaristo, T. E., Korpi-Hyövälti, E., Saltevo, J., Laakso, M., Kuusisto, J., Metspalu, A., Collins, F. S., Mohlke, K. L., Bergman, R. N., Tuomilehto, J., Boehm, B. O., Gieger, C., Hveem, K., Cauchi, S., Froguel, P., Baldassarre, D., Tremoli, E., Humphries, S. E., Saleheen, D., Danesh, J., Ingelsson, E., Ripatti, S., Salomaa, V., Erbel, R., Jöckel, K. H., Moebus, S., Peters, A., Illig, T., de Faire, U., Hamsten, A., Morris, A. D., Donnelly, P. J., Frayling, T. M., Hattersley, A. T., Boerwinkle, E., Melander, O., Kathiresan, S., Nilsson, P. M., Deloukas, P., Thorsteinsdottir, U., Groop, L. C., Stefansson, K., Hu, F., Pankow, J. S., Dupuis, J., Meigs, B., Altshuler, D., Boehnke, M., McCarthy, M. I. and DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.*, **44**, 981–990.
Niemi, J. (2010) Evaluating individual player contributions in basketball. *Proc. Statist. Comput. Sect. Am. Statist. Ass.*, 4914–4923.
Noma, H., Matsui, S., Omori, T. and Sato, T. (2010) Bayesian ranking and selection methods using hierarchical mixture models in microarray studies. *Biostatistics*, **11**, 281–289.
Normand, S.-L. T., Glickman, M. E. and Gatsonis, C. A. (1997) Statistical methods for profiling providers of medical care: issues and applications. *J. Am. Statist. Ass.*, **92**, 803–814.
Paddock, S. M. and Louis, T. A. (2011) Percentile-based empirical distribution function estimates for performance evaluation of healthcare providers. *Appl. Statist.*, **60**, 575–589.
Pyeon, D., Newton, M. A., Lambert, P. F., Den Boon, J. A., Sengupta, S., Marsit, C. J., Woodworth, C. D., Connor, J. P., Haugen, T. H., Smith, E. M., Kelsey, K. T., Turek, L. P. and Ahlquist, P. (2007) Fundamental differences

in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. *Cancer Res.*, **67**, 4605–4619.
Richey, M. and Zorn, P. (2005) Basketball, beta, and Bayes. *Math. Mag.*, **78**, 354–367.
Shen, W. and Louis, T. A. (1998) Triple-goal estimates in two-stage hierarchical models. *J. R. Statist. Soc.* B, **60**, 455–471.
Smyth, G. K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statist. Appl. Genet. Mol. Biol.*, **3**, article 3.
Storey, J. D. (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.*, **31**, 2013–2035.
van der Vaart, A. (1998) *Asymptotic Statistics.* Cambridge: Cambridge University Press.
Wright, D. L., Stern, H. S. and Cressie, N. (2003) Loss functions for estimation of extrema with an application to disease mapping. *Can. J. Statist.*, **31**, 251–266.
Xie, M., Singh, K. and Zhang, C.-H. (2009) Confidence intervals for population ranks in the presence of ties and near ties. *J. Am. Statist. Ass.*, **104**, 775–788.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material'.