Statistical Applications in Genetics and Molecular Biology

Volume 6, *Issue* 1 2007 *Article* 9

Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach

Rainer Opgen-Rhein* Korbinian Strimmer†

*Department of Statistics, University of Munich, rainer.opgen-rhein@stat.uni-muenchen.de

†Department of Statistics, University of Munich, strimmer@uni-leipzig.de

Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach*

Rainer Opgen-Rhein and Korbinian Strimmer

Abstract

High-dimensional case-control analysis is encountered in many different settings in genomics. In order to rank genes accordingly, many different scores have been proposed, ranging from ad hoc modifications of the ordinary t statistic to complicated hierarchical Bayesian models.

Here, we introduce the "shrinkage t" statistic that is based on a novel and model-free shrinkage estimate of the variance vector across genes. This is derived in a quasi-empirical Bayes setting. The new rank score is fully automatic and requires no specification of parameters or distributions. It is computationally inexpensive and can be written analytically in closed form.

Using a series of synthetic and three real expression data we studied the quality of gene rankings produced by the "shrinkage t" statistic. The new score consistently leads to highly accurate rankings for the complete range of investigated data sets and all considered scenarios for across-gene variance structures.

KEYWORDS: high-dimensional case-control data, James-Stein shrinkage, limited-translation, quasi-empirical Bayes, regularized t statistic, variance shrinkage

---

*Acknowledgments: This research was supported by an Emmy Noether excellence grant from the Deutsche Forschungsgemeinschaft. We thank Gary Churchill and an anonymous referee for valuable comments.

Introduction

High-dimensional case-control analysis is a key problem that has many applications in computational genomics. The most well-known example is that of ranking genes according to differential expression, but there are many other instances that warrant similar statistical methodology, such as the problem of detecting peaks in mass spectrometric data or finding genomic enrichment sites.

All these problems have in common that they require a variant of a (regularized) t statistic that is suitable for high-dimensional data and large-scale multiple testing. For this purpose, in the last few years various test statistics have been suggested, which may be classified as follows:

i) Simple methods: fold change, classical t statistic.

ii) Ad hoc modifications of ordinary t statistic: Efron's 90% rule (Efron et al., 2001), SAM (Tusher et al., 2001).

iii) (Penalized) likelihood methods, e.g.: Ideker et al. (2000), Wright and Simon (2003), Wu (2005).

iv) Hierarchical Bayes methods, e.g.: Newton et al. (2001), Baldi and Long (2001), Lönnstedt and Speed (2002), Newton et al. (2004), "moderated t" (Smyth, 2004), Cui et al. (2005), Fox and Dimmic (2006).

For an introductory review of most of these approaches see, e.g., Cui and Churchill (2003) and Smyth (2004).

Current good practice in gene expression case-control analysis favors the empirical or full Bayesian approaches (item iv) over other competing methods. The reason behind this is that Bayesian methods naturally allow for information sharing across genes, which is essential when the number of sample is as small in typical genomic experiments. Specifically, the estimation of gene-specific variances profits substantially from pooling information across genes (e.g., Wright and Simon, 2003; Smyth, 2004; Delmar et al., 2005; Cui et al., 2005). On the other hand, Bayesian methods can become computationally quite expensive, and more importantly, typically rely on a host of very detailed assumptions concerning the underlying data and parameter generating models.

In this paper we introduce a novel "shrinkage t" approach that is as simple as the ad hoc rules (item ii) but performs as well as fully Bayesian models (item iv), even in simulation settings that are favorable to the latter. Moreover, the new gene ranking statistic is fully analytic, requires no computer-intensive procedures, and is derived without any specific distributional assumptions. In this sense, it is a

further development of the quasi-likelihood approach of Strimmer (2003) but with additional regularization.

The shrinkage *t* statistic is developed in the framework of James-Stein-type analytic shrinkage (e.g., Gruber, 1998; Schäfer and Strimmer, 2005). This approach offers a highly efficient means for regularized inference, both in the statistical and computational sense. It is complementary to more well-known alternatives such as Bayesian and penalized likelihood inference. Nevertheless, the resulting estimators are typically very hard to improve (Yi-Shi Shao and Strawderman, 1994). James-Stein shrinkage estimation may also be understood as a "quasi-empirical Bayes" method as only information concerning second moments rather than fully specified distributions are used. In short, analytic shrinkage estimators combine properties that render them very attractive for analyzing large-dimensional genomic assays.

In the context of differential expression a similar approach was suggested before only by Cui et al. (2005) who also employ James-Stein estimation to obtain shrinkage estimates of the gene-specific variances. Our approach shares many aspects with that of Cui et al. (2005). However, our estimator for variance shrinking is different in that absolutely no distributional assumptions are involved (not even for hyperparameters). Moreover, it is applied on the original data scale and thus requires no transformations. Finally, it is derived via a rather general route for constructing Stein-type estimators, and results in a very compact, fully analytic, and yet still highly efficient estimator for variance shrinkage (Eq. 10 and Eq. 11).

The remainder of this paper is structured as follows. In the next section we briefly review analytic shrinkage estimation. Subsequently, we develop an estimator for inference of gene-specific variances and construct the shrinkage *t* score for ranking differentially expressed genes. Subsequently, we investigate the performance of this statistic in simulations and in extensive data analysis in comparison relative to a number of competing statistics. The final section contains a discussion of the results.

Distribution-Free Shrinkage Estimation

In this section we describe how analytic James-Stein-type shrinkage estimators may be constructed from an arbitrary unregularized estimator, without assuming any distributions for data or the model parameters.

James-Stein Shrinkage Rules

Initially, we assume that an unregularized estimation rule

$$\delta^0 = \hat{\boldsymbol{\theta}}, \tag{1}$$

is available, e.g., the maximum-likelihood or the minimum variance unbiased estimate. It is important here that $\hat{\boldsymbol{\theta}}$ is a *vector* $(\theta_1, \ldots, \theta_k, \ldots, \theta_p)^T$. (In the specific example of the present article this vector contains all gene-specific empirical variances.) Then the James-Stein ensemble shrinkage estimation rule may be written as

$$\begin{aligned} \delta^\lambda &= \delta^0 - \lambda \boldsymbol{\Delta} \\ &= \hat{\boldsymbol{\theta}} - \lambda \, (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{\text{Target}}). \end{aligned} \tag{2}$$

In other words, the shrinkage estimate δ^λ is the linear combination $\lambda\hat{\boldsymbol{\theta}}^{\text{Target}} + (1-\lambda)\hat{\boldsymbol{\theta}}$ of the original estimator $\hat{\boldsymbol{\theta}}$ and a target estimate $\hat{\boldsymbol{\theta}}^{\text{Target}}$. The parameter λ determines the extent to which these estimates are pooled together. If $\lambda = 1$ then the target dominates completely, whereas for $\lambda = 0$ no shrinkage occurs.

In James-Stein estimation the search for the optimal shrinkage intensity λ is considered from a decision theoretic perspective. First, a loss function is selected (e.g. the squared error). Second, λ is chosen such that the corresponding risk of δ^λ, i.e. the expectation of the loss with respect to the data (e.g. the mean squared error, MSE), is minimized.

Interestingly, it turns out that for squared error loss this can be done *without* any reference to the unknown true value $\boldsymbol{\theta}$, as the MSE of δ^λ may be written as follows:

$$\begin{aligned} \text{MSE}(\delta^\lambda) = \text{MSE}(\hat{\boldsymbol{\theta}}) &+ \lambda^2 \sum_{k=1}^{p} \{E((\hat{\theta}_k - \hat{\theta}_k^{\text{Target}})^2)\} \\ &- 2\lambda \sum_{k=1}^{p} \{\text{Var}(\hat{\theta}_k) - \text{Cov}(\hat{\theta}_k, \hat{\theta}_k^{\text{Target}}) \\ &\qquad + \text{Bias}(\hat{\theta}_k)\, E(\hat{\theta}_k - \hat{\theta}_k^{\text{Target}})\} \\ =:\ c &+ \lambda^2\, b - 2\lambda\, a. \end{aligned} \tag{3}$$

Hence, the MSE risk curve has the shape of a parabola whose parameters a, b, and c are completely determined by only the first two distributional moments of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^{\text{Target}}$. This allows a number of further insights concerning the shrinkage rule Eq. 2:

- The risk improvement of δ^{λ} compared to MSE of the unregularized estimate δ^0 is determined only by a and b (note that $\text{MSE}(\delta^0) = c$).
- *Any* value of λ in the range between 0 and $2\frac{a}{b}$ leads to a decrease in MSE.
- The optimal shrinkage intensity that results in overall minimum MSE is given by the simple formula
$$\lambda^{\star} = \frac{a}{b}. \tag{4}$$
- In this case the savings relative to the unshrunken estimate amount to $\text{MSE}(\hat{\boldsymbol{\theta}}) - \text{MSE}(\delta^{\lambda^{\star}}) = \frac{a^2}{b}$.
- The factor b, which measures the misspecification of target and estimate, plays the role of a precision for λ.

Further discussion and interpretation of Eq. 4 may be found in Schäfer and Strimmer (2005). We only note here that special versions of the rule $\lambda^{\star} = \frac{a}{b}$ are well known, see, e.g., Ledoit and Wolf (2003) who describe the multivariate case but require an unbiased $\hat{\boldsymbol{\theta}}$, or Thompson (1968) who considers only the univariate case and a non-stochastic target and also restricts to $\text{Bias}(\hat{\theta}) = 0$.

Construction of Shrinkage Estimator

In actual application of the shrinkage rule (Eq. 2) the pooling parameter λ needs to be estimated from the data. Inevitably, this leads to an *increase* of the total risk of the resulting shrinkage estimator. However, it is a classic result by Stein that the cost of estimating the shrinkage intensity is already (and always!) offset by the savings in total risk when the dimension p is larger than three (e.g. Gruber, 1998).

One straightforward way to estimate the optimal $\lambda^{\star}$ is to replace the variances and covariances in Eq. 3 by their unbiased empirical counterparts, i.e. $\hat{\lambda}^{\star} = \frac{\hat{a}}{\hat{b}}$. Alternatively, an unbiased estimate for the whole fraction $\frac{a}{b}$ may be sought – but this will only be possible in some special cases. We would also like to point out that we do *not* recommend the suggestion of Thompson (1968) who employ a biased estimate for the denominator (b) to Eq. 4 – this will lead to a potentially very inaccurate shrinkage estimator.

Despite its simplicity, rule Eq. 2 together with an estimated version of Eq. 4 provides instant access to several classic shrinkage estimators, and offers a simple and unified framework for their derivation.

For instance, consider the old problem of Stein (1956, 1981) of inferring the mean of a p-dimensional multivariate normal distribution with unit-diagonal covariance matrix from a single ($n = 1$) vector-valued observation – clearly an extreme example of the "small n, large p" setting. In this case the maximum-likelihood estimate equals the vector of observations, i.e. $\hat{\theta}_k^{\mathrm{ML}} = x_k$. However, shrinkage estimators with improved efficiency over the ML estimator are easily constructed. With the covariance being the identity matrix ($\mathrm{Var}(x_k) = 1$ and $\mathrm{Cov}(x_k, x_l) = 0$) and the target being set to zero ($\hat{\boldsymbol{\theta}}^{\mathrm{Target}} = 0$) one finds $a = p$ and $\hat{b} = \sum_{k=1}^{p} x_k^2$ which results in the shrinkage estimator

$$\hat{\theta}_k^{\mathrm{JS}} = \left(1 - \frac{p}{\sum_{k=1}^{p} x_k^2}\right) x_k. \tag{5}$$

If we follow Lindley and Smith (1972) and shrink instead towards the mean across dimensions $\bar{x} = \frac{1}{p}\sum_{k=1}^{p} x_k$ we get $a = p - 1$ and $\hat{b} = \sum_{k=1}^{p}(x_k - \bar{x})^2$ and obtain

$$\hat{\theta}_k^{\mathrm{EM}} = \bar{x} + \left(1 - \frac{p-1}{\sum_{k=1}^{p}(x_k - \bar{x})^2}\right)(x_k - \bar{x}) \tag{6}$$

It is noteworthy that these are *not* the original Stein estimators given in James and Stein (1961) and Efron and Morris (1973) but instead are exactly the shrinkage estimators of Stigler (1990) derived using a regression approach. We point out that the Stigler and our versions have the advantage that they are applicable also for $p = 1$ and $p = 2$.

Positive Part Estimator and Component Risk Protection by Limited Translation

The efficiency of the above shrinkage estimator can be further improved by two simple measures.

Firstly, by truncating the estimated $\hat{\lambda}$ at one,

$$\delta^{\hat{\lambda}+} = \delta^0 - \min(1, \hat{\lambda})\boldsymbol{\Delta}, \tag{7}$$

which results in the so-called positive part James-Stein estimator that dominates the unrestricted shrinkage estimator of Eq. 2 in terms of statistical efficiency (Barachnik, 1970).

Secondly, by restricting the translation allowed for individual components. The original James-Stein procedure is geared, in the terminology of Efron and Morris (1975), towards producing estimators with good *ensemble* risk properties. This

means that it aims at minimizing the total risk accumulated over all parameters. However, in some instances this may occur at the expense of individual parameters whose risks may even increase (!). Therefore, in Stein estimation (and indeed also in hierarchical Bayes estimation) individual components of a parameter vector need to be protected against too much shrinkage.

"Limited translation" (Efron and Morris, 1972, 1975) is a simple way to construct estimators that exhibit both good ensemble risk as well as favorable component risk properties. One example of a protected shrinkage rule is

$$\delta_k^{\hat{\lambda}+,M} = \delta_k^0 - \min(1, \hat{\lambda}) \min(1, \frac{M}{|\Delta_k|})\Delta_k, \tag{8}$$

which ensures that we always have $|\delta_k^{\hat{\lambda}+,M} - \delta_k^0| \leq M$, where M is a cutoff parameter chosen by the user. A convenient selection of M is, e.g., the 99 percent quantile of the distribution of the absolute values $|\Delta_k|$ of the components of the shrinkage vector $\mathbf{\Delta}$. In the terminology of Efron and Morris (1972), the term $\min(1, \frac{M}{|\Delta_k|})$ constitutes the *relevance function* that determines the degree to which any particular component is affected by the ensemble-wide shrinkage.

Finally, we point out an interesting connection with soft thresholding, as the above limited translation shrinkage rule may also be written as

$$\delta_k^{\hat{\lambda}+,M} = \delta_k^{\hat{\lambda}+} + \min(1, \hat{\lambda})(|\Delta_k| - M)_+ \mathrm{sgn}(\Delta_k), \tag{9}$$

where the subscript "+" denotes truncation at zero.

Further Remarks

In order to complete the discussion of analytic James-Stein shrinkage estimators we would like to remark on the following additional points:

- It is interesting to note that many empirical Bayes estimators can be put into the form of Eq. 2, (e.g. Gruber, 1998). Note that using Eq. 3 allows to derive these estimators without first going through the full Bayesian formalism!
- Using the above equation leads to an almost automatic procedure for shrinkage estimation.
- The construction of the estimator assumes at no point a normal or any other distribution.
- Note that it is possible to allow for multiple shrinkage intensities. For instance, if the model parameters fall into two natural groups, each could have

its own target and its own associated shrinkage intensity. In the extreme case each parameter could have its own λ.

The "Shrinkage *t*" Statistic

Shrinkage Estimation of Variance Vector

Within the above framework for distribution-free shrinkage it is straightforward to construct an efficient estimator of gene-specific variances.

From given data with p variables (genes) we first compute the usual unbiased empirical variances $v_1^2, \ldots, v_p^2$. These provide the components for the unregularized vector estimate $\hat{\boldsymbol{\theta}}$ of Eq. 1. Subsequently, we choose a suitable shrinkage target. For this we suggest using the median value of all v_k. In the exploration of possible other targets we considered also shrinking against zero and towards the mean of the empirical variances. However, these two alternatives turned out to be either less efficient (zero target) or less robust (mean target) than shrinking towards the median.

Following the recipe outlined above, we immediately obtain the shrinkage estimator

$$v_k^\star = \hat{\lambda}^\star v_{\text{median}} + (1 - \hat{\lambda}^\star) v_k \tag{10}$$

with optimal estimated pooling parameter

$$\hat{\lambda}^\star = \min\left(1, \frac{\sum_{k=1}^p \widehat{\text{Var}}(v_k)}{\sum_{k=1}^p (v_k - v_{\text{median}})^2}\right), \tag{11}$$

Note that in this formula we have used the approximation $\text{Cov}(v_k, v_{\text{median}}) \approx 0$.

Eq. 11 has an intuitive interpretation. If the empirical variances v_k can be reliably determined from the data, and consequently exhibit only a small variance themselves, there will be little shrinkage, whereas if $\widehat{\text{Var}}(v_k)$ is comparatively large pooling across genes will take place. Furthermore, the denominator of Eq. 11 is an estimate of the misspecification between the target and the v_k. Hence, if the target is incorrectly chosen then no shrinkage will take place either.

The computation of a sample version of $\widehat{\text{Var}}(v_k)$ is straightforward. Defining $\bar{x}_k = \frac{1}{n}\sum_{i=1}^n x_{ik}$, $w_{ik} = (x_{ik} - \bar{x}_k)^2$, and $\bar{w}_k = \frac{1}{n}\sum_{i=1}^n w_{ik}$, we have $v_k = \frac{n}{n-1}\bar{w}_k$ and $\widehat{\text{Var}}(v_k) = \frac{n}{(n-1)^3}\sum_{i=1}^n (w_{ik} - \bar{w}_k)^2$. A similar formula can be derived for the variance of the entries of the empirical covariance matrix (cf. Schäfer and Strimmer (2005)).

Construction of "Shrinkage *t*" Statistic

The "shrinkage *t*" statistic considered in the following is obtained by plugging the above shrinkage variance estimate (Eq. 10 and Eq. 11) into the ordinary *t* statistic. With the sample sizes in groups 1 and 2 denoted as n_1 and n_2 the shrinkage *t* statistic is given by

$$t_k^\star = \frac{\bar{x}_{k1} - \bar{x}_{k2}}{\sqrt{\frac{v_{k1}^\star}{n_1} + \frac{v_{k2}^\star}{n_2}}}. \tag{12}$$

We consider two variants of this statistic, one where variances are estimated separately in each group (hence with two different shrinkage intensities), and the other one using a pooled estimate (i.e. with one common shrinkage factor).

Note that the shrinkage *t* statistic essentially provides a compromise between the standard *t* statistic (to which it reduces for $\lambda = 0$) and the fold change or difference of means statistic ($\lambda = 1$).

Other Regularized *t* Statistics

Closely related to the shrinkage *t* statistic are in particular two approaches, "moderated *t*" (Smyth, 2004) and the Stein-type procedure by Cui et al. (2005). The essential feature characteristic for both of these methods is, again, variance shrinkage, albeit done in a different fashion compared to shrinkage *t*:

1. The moderated *t* method assumes a scale-inverse-chi-square distribution as distribution for the variances across genes. The corresponding parameters are estimated by empirical Bayes, and the resulting variance estimates are plugged into the *t*-statistic.

2. The variance shrinking procedure of Cui et al. (2005) is essentially the classic Stein estimator on the log-scale, complemented by a bias correction and by an estimation of the variance factor (numerator in the Stein formula) by simulation from a chi-square distribution whose degrees of freedom depends on the sample size. The *t* statistic resulting from using this kind of variance shrinkage will be called "Cui et al. *t*" in this paper.

Therefore, both the moderated *t* and the Cui et al. *t* rely on some form of distributional assumption, whereas the shrinkage *t* statistic is derived without any such consideration.

Results

Assessment of Quality of Gene Ranking

In this section we describe the results from computer simulations and analysis of experimental gene expression data that we conducted to assess the quality of gene ranking provided by the shrinkage t statistic in comparison to other competing scores.

Setup of Simulations

In the simulations we followed closely the setup specified in Smyth (2004):

- Variances across genes were assumed to follow a scale-inverse-chi-square distribution Scale-inv-$\chi^2(d_0, s_0^2)$ with $s_0^2 = 4$ and three different settings for the degrees of freedom d_0: highly similar variances across genes ($d_0 = 1000$), balanced variances ($d_0 = 4$), and different variances across genes ($d_0 = 1$).
- In total 2,000 genes were considered, 100 of which were randomly assigned to be differentially expressed.
- The differences in group means for the 100 differentially expressed genes were determined by drawing from a Normal distribution with mean zero and the gene-specific variance, whereas for the non-differentially expressed genes it was set to zero.
- Finally, synthetic data matrices were obtained by sampling for each gene and separately for the control and case groups three independent observations from a Normal distribution with the respective gene-specific variances and means.

These data formed the basis for computing various gene ranking scores. Specifically, we compared the following statistics: fold change, ordinary t, moderated t (Smyth, 2004), Cui et al. t statistic (i.e. the unequal variance t statistic regularized by using the variances estimated by the method of Cui et al. (2005)), Efron's 90% rule (Efron et al., 2001), Wu's improved SAM statistic (Wu, 2005), and the shrinkage t statistic (with both equal and unequal variances). As reference we also included random ordering in the analysis. For these different ways of producing rankings we computed false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN) for all possible cut-offs in the gene list (1-2000).

This procedure was repeated 500 times for each test statistic and variance scenario, to obtain estimates of the true discovery rates $E(\frac{TP}{TP+FP})$ and ROC curves describing the dependence of sensitivity $E(\frac{TP}{TP+FN})$ and specificity $E(\frac{TN}{TN+FP})$.

Experimental Data Sets

In addition to the simulations we computed the gene ranking for three experimental case-control data sets with known differentially expressed genes.

The first data set studied is the well-known Affymetrix spike-in study that contains 12,626 genes, 12 replicates in each group, and 16 known differentially expressed genes (Cope et al., 2004).

The second investigated data is a subset of the "golden spike" Affymetrix experiment of Choe et al. (2005). From the original data we removed the 2,535 probe sets for spike-ins with ratio 1:1, leaving in total 11,475 genes with 3 replicates per group, and 1,331 known differentially expressed genes. We note that excluding the 1:1 spike-ins is important as these comprise a severe experimental artifact (Irizarry et al., 2006). Both the Choe et al. (2005) data and the Affymetrix spike-in data were calibrated and normalized using the default methods of the "affy" R package (Gautier et al., 2004).

The third data set is from the HIV-1 infection study of van 't Wout et al. (2003). It contains 4 replicates per group, and 13 of the 4,608 genes have been experimentally confirmed to be differentially expressed.

For reproducibility, these three experimental test data sets are available for download from `http://strimmerlab.org/data.html` exactly in the form used in this article, including all preprocessing.

Performance of Gene Ranking Statistics

The results from simulations and data analysis are summarized in Fig. 1 and Fig. 2. In each figure the first row shows the fraction of correctly identified differentially expressed genes in relation to the number of included genes (i.e. the true discovery rate, or positive predictive value), whereas the second row depicts the corresponding receiver-operator characteristic (ROC).

If the variances are highly similar across genes (Fig. 1, first column) the best methods are fold change, moderated t, Efron t, Cui et al. t, and shrinkage t, all of which provide in this case similarly accurate rankings. Only the ordinary t statistics and Wu's improved SAM statistic are not efficient in this setting. If variances are balanced (Fig. 1, second column), the best rankings are given by moderated t, Efron t and shrinkage t. The gene ranking accuracy of fold-change is dropping to that of the standard t statistic. If the variances are highly different (Fig. 1, column 3), fold change ranking is close to random, and Efron's 90% percent rule also becomes inefficient. Only moderated t and shrinkage t are offering optimal gene rankings in this setting. The Cui et al. t and the ordinary t test produce similar and the second best rankings in this case.

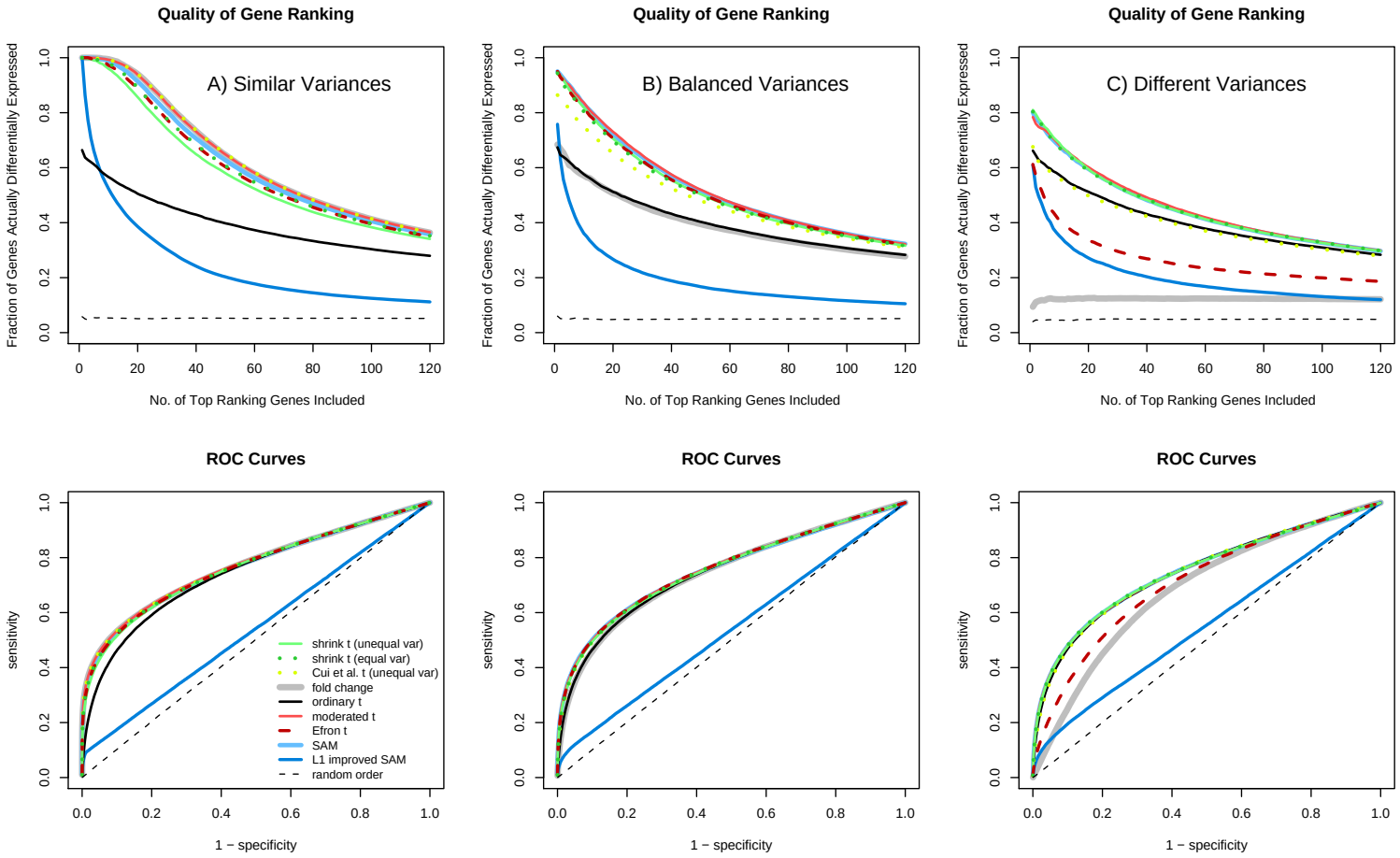

Figure 1: True discovery rates and ROC curves computed for simulated data under three different scenarios for the distribution of variances across genes. See main text for details of simulations and analysis procedures.

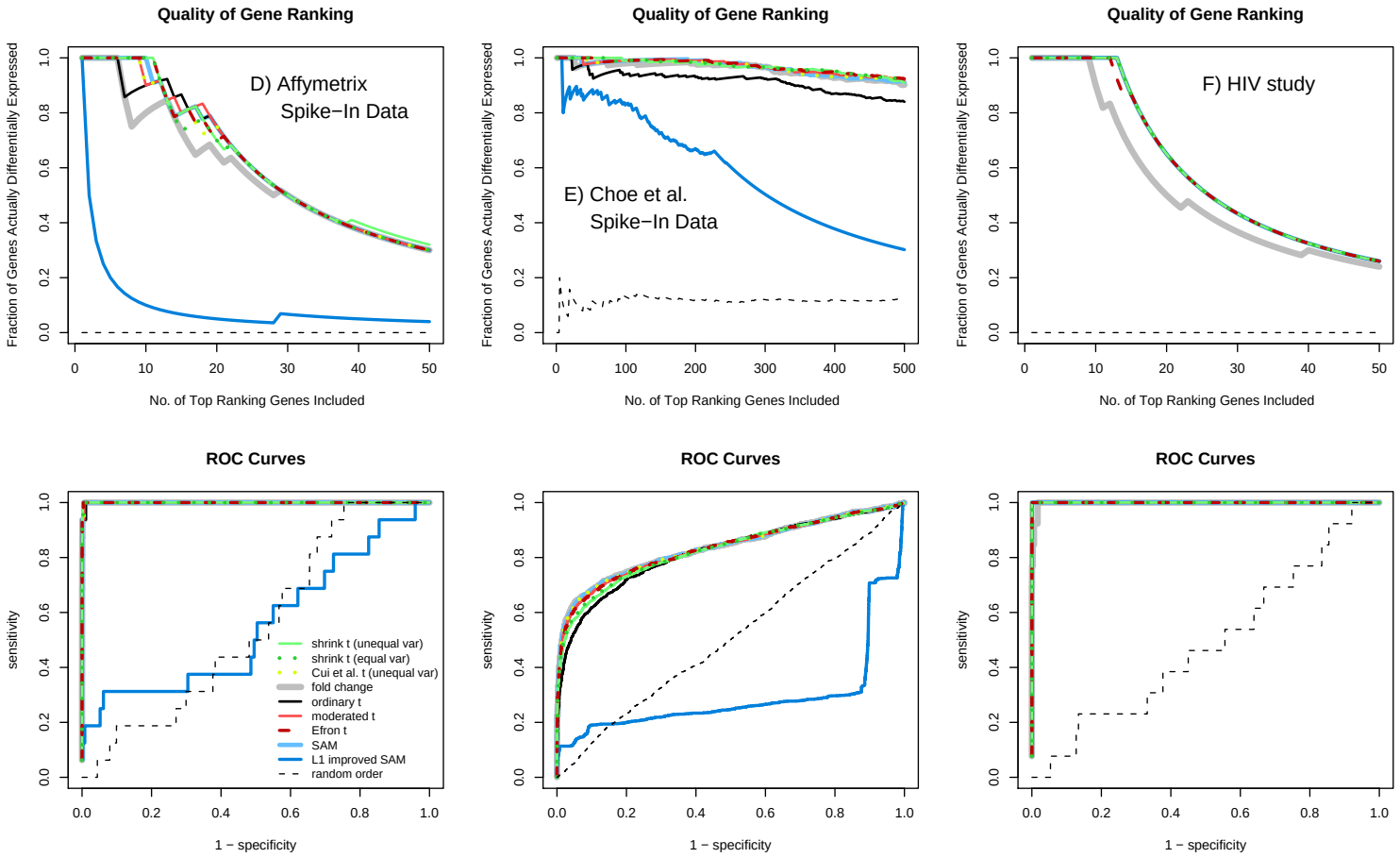

Figure 2: True discovery rates and ROC curves for the three investigated experimental data sets.

In the analysis of the Affymetrix spike-in data the methods with largest true discovery rates independent of the chosen cutoff are the shrinkage *t* statistic (unequal variance), Efron *t*, shrinkage *t* statistic (equal variance), moderated *t*, and Cui et al. *t*. For the Choe et al. (2005) data the shrinkage *t* statistic shows the best performance along with moderated *t*, Efron *t*, Cui et al. *t*, and fold change, whereas the ordinary *t* statistic and the improved SAM statistic don't perform well. Thus, this data set resembles the situation in the simulations where all variances were highly similar. Finally, for the van 't Wout et al. (2003) study all methods except for fold-change provide optimal rankings, with the Efron *t* statistic being slightly less accurate than the remainder of the methods.

This generally confirms earlier findings presented in Smyth (2004). We emphasize in addition the following points:

- The ordinary *t* statistic shows average though never optimal performance regardless of the variance structure across genes.
- Using fold change is only a good idea if variances are all fairly similar; the same is true for Efron *t* statistic.
- The improved SAM statistic by Wu (2005) generally provides very poor rankings of genes. This is due to the fact that most genes with differential expression are assigned a zero score, and hence are ordered randomly (i.e. in input order).
- ROC curves appear to be only of limited help for assessing performance. Instead, we suggest relying on quality of ranking plots based on estimated true discovery rates.
- The shrinkage *t* and the moderated *t* statistics are the only two methods that perform optimally in all three simulation settings. However, the Cui et al. *t* statistic is nearly as good, showing only a slightly decreased accuracy when variances are strongly heterogeneous.
- Both moderated *t* and shrinkage *t* produce accurate rankings also for the three experimental data, perhaps with a small edge for shrinkage *t* in the Affymetrix spike-in study.
- The shrinkage *t* statistic with unequal variances performs nearly as well as shrinkage *t* with equal variances, even though the former has only half the sample size available for estimation of variances.

We also note that the moderated *t* statistic is based exactly on the model employed in the simulations (i.e. scale-inverse-chi-square distribution for the variance). The Cui

et al. *t* statistic also incorporates prior knowledge on the distribution of variances across genes in its simulation step employing a chi-square distribution.

Therefore, it is easy to understand why moderated *t* and the Cui et al. *t* statistic perform well. On the other hand, we point out that it is quite remarkable that shrinkage *t*, a simple analytic statistic not specifically tailored to any particular distributional setting, can fully match the performance of the moderated *t* approach.

Discussion

In this paper we have introduced a novel gene ranking statistic for genomic case control studies. This method is based on a James-Stein-type shrinkage approach that, unlike its Bayesian or empirical Bayes cousins, is fully analytic and does not rely on explicit priors or other distributional assumptions. Hence, this approach is potentially more flexible, for instance when variance scenarios differ (e.g. Gelman, 2006), the underlying models are misspecified, or when variances are unequal. Most importantly, the proposed method provides highly accurate gene rankings both for simulated and real data, on par with much more complicated models, but without relying on computational expensive procedures such as MCMC or optimization.

From our simulations it is also interesting to learn that there seems to exist an optimality limit with regard to producing accurate rankings. In our comparative evaluation the moderated *t* statistic and the shrinkage *t* statistic were the only two methods that achieved that limit for all considered scenarios.

In this paper, we haven't raised at all the issue of (multiple) testing needed to determine an appropriate cut-off value. This is typically done by controlling the false discovery rate. Our preferred tools for this task are mixture models (e.g., Sapir and Churchill, 2000; Dean and Raftery, 2005) and the "local fdr" approach (see, e.g. Efron, 2005). The latter procedure has the advantage of being adaptive with regard to the null hypothesis. This means that it will automatically take account of correlation among the genes, and also accommodate for the decreased variance of the null-distribution of the shrinkage *t* statistic, which by construction is smaller than that of the standard *t* statistic. However, with Fig. 1 in mind we caution that it often is not possible to guarantee a prescribed false discovery rate even in optimal circumstances - as this crucially depends on the capability of the underlying statistic to produce an accurate ranking!

In summary, with few exceptions (e.g., Cui et al., 2005; Schäfer and Strimmer, 2005) James-Stein-type estimation appears to have been somewhat overlooked in the recent efforts for analyzing high-dimensional systems (it is not mentioned in the reference text by Hastie et al. (2001), for instance). In this respect, the shrinkage

t approach demonstrates that statistics derived in this fashion may indeed compare very favorable to penalized ML or Bayesian methods. Indeed, our proposed variance shrinkage procedure may be useful not only in simple *t* test situations but also in more general ANOVA-type analyses (Smyth, 2004; Cui et al., 2005).

A Computer Implementation and Availability

All statistical procedures described have been implemented in computer programs that are available under the terms of the GNU General Public License.

The "shrinkage *t*" statistic is implemented in the R package "st" which is available from the CRAN archive (`http://cran.r-project.org`) and from web page `http://strimmerlab.org/software/st/`. This package also contains wrapper functions for a number of other regularized *t* statistics.

The shrinkage variance estimator of Eq. 10 and Eq. 11 is contained in the R package "corpcor" that is available from `http://strimmerlab.org/software/corpcor/` and also from CRAN.

Note added in proof:

Since writing of this paper a Stein-type approach to variance shrinkage similar to ours has appeared in Tong and Wang (2007).

References

Baldi, P. and A. D. Long (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics 17*, 509–519.

Barachnik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Statist. 41*, 642–645.

Choe, S. E., M. Boutros, A. M. Michelson, G. M. Church, and M. S. Halfon (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control data set. *Genome Biology 6*, R16.

Cope, L. M., R. A. Irizarray, H. A. Jaffee, Z. Wu, and T. P. Speed (2004). A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics 20*, 323–331.

Cui, X. and G. A. Churchill (2003). Statistical test for differential expression in cDNA microarray experiments. *Genome Biology 4*, R210.

Cui, X., J. T. G. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics 6*, 59–75.

Dean, N. and A. E. Raftery (2005). Normal uniform mixture differential gene expression detection for cDNA microarrays. *Genome Biology 6*, 173.

Delmar, P., S. Robin, D. Tronik-Le Roux, and J. J. Daudin (2005). Mixture model on the variance for the differential analysis of gene expression data. *Appl. Statist. 54*, 31–50.

Efron, B. (2005). Local false discovery rates. Technical report, Department of Statistics, Stanford University.

Efron, B. and C. N. Morris (1972). Limiting the risk of Bayes and empirical Bayes estimators – part II: The empirical Bayes case. *J. Amer. Statist. Assoc. 67*, 130–139.

Efron, B. and C. N. Morris (1973). Stein's estimation rule and its competitors–an empirical Bayes approach. *J. Amer. Statist. Assoc. 68*, 117–130.

Efron, B. and C. N. Morris (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc. 70*, 311–319.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc. 96*, 1151–1160.

Fox, R. J. and M. W. Dimmic (2006). A two-sample Bayesian t-test for microarray data. *BMC Bioinformatics 7*, 126.

Gautier, L., L. Cope, B. M. Bolstad, and R. A. Irizarry (2004). affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics 20*, 307–315.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*, 515–533.

Gruber, M. H. J. (1998). *Improving Efficiency By Shrinkage*. New York: Marcel Dekker, Inc.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer Verlag.

Ideker, T., V. Thorsson, A. F. Seigel, and L. E. Hood (2000). Testing for differentially expressed genes by maximum likelihood analysis of microarray data. *J. Comp. Biol. 7*, 805–817.

Irizarry, R. A., L. Cope, and Z. Wu (2006). Feature-level exploration of a published control data set. *Genome Biology 7*, 404.

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, Volume 1, Berkeley, pp. 361–379. Univ. California Press.

Ledoit, O. and M. Wolf (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finance 10*, 603–621.

Lindley, D. V. and A. F. M. Smith (1972). Bayes estimates for the linear model. *J. R. Statist. Soc. B 34*, 1–72.

Lönnstedt, I. and T. Speed (2002). Replicated microarray data. *Statistica Sinica 12*, 31–46.

Newton, M. A., C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comp. Biol. 8*, 37–52.

Newton, M. A., A. Noueiry, D. Sarkar, and P. Ahlquist (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics 5*, 155–176.

Sapir, M. and G. A. Churchill (2000). Estimating the posterior probability of differential gene expression from microarray data. Poster, Jackson Laboratory, Bar Harbor.

Schäfer, J. and K. Strimmer (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol. 4*, 32.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statist. Appl. Genet. Mol. Biol. 3*, 3.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In J. Neyman (Ed.), *Proc. Third Berkeley Symp. Math. Statist. Probab.*, Volume 1, Berkeley, pp. 197–206. Univ. California Press.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist. 6*, 1135–1151.

Stigler, S. M. (1990). A Galtonian perspective on shrinkage estimators. *Statistical Science 5*, 147–155.

Strimmer, K. (2003). Modeling gene expression measurement error: a quasi-likelihood approach. *BMC Bioinformatics 4*, 10.

Thompson, J. R. (1968). Some shrinkage techniques for estimating the mean. *J. Amer. Statist. Assoc. 63*, 113–122.

Tong, T. and Y. Wang (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis. *J. Amer. Statist. Assoc. 102*, 113–122.

Tusher, V., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA 98*, 5116–5121.

van 't Wout, A. B., G. K. Lehrman, S. A. Mikheeva, G. C. O'Keeffe, M. G. Katze, R. E. Bumgarner, G. K. Geiss, and J. I. Mullins (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4(+)-T-cell lines. *J Virol 77*(2), 1392–1402.

Wright, G. W. and R. M. Simon (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics 19*, 2448–2455.

Wu, B. (2005). Differential gene expression detection using penalized linear regression models: the improved SAM statistic. *Bioinformatics 21*, 1565–1571.

Yi-Shi Shao, P. and W. E. Strawderman (1994). Improving on the James-Stein positive-part estimator. *Ann. Statist. 22*, 1517–1538.