Biostatistics Advance Access published May 5, 2016

Biostatistics (2016), *pp.* 1–15
doi:10.1093/biostatistics/kxw018

Study design in high-dimensional classification analysis

BRISA N. SÁNCHEZ*
Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA
brisa@umich.edu

MEIHUA WU
Gilead Sciences, Inc, Foster City, CA 94404, USA

PETER X. K. SONG, WEN WANG
Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

Summary

Advances in high throughput technology have accelerated the use of hundreds to millions of biomarkers to construct classifiers that partition patients into different clinical conditions. Prior to classifier development in actual studies, a critical need is to determine the sample size required to reach a specified classification precision. We develop a systematic approach for sample size determination in high-dimensional (large p small n) classification analysis. Our method utilizes the probability of correct classification (PCC) as the optimization objective function and incorporates the higher criticism thresholding procedure for classifier development. Further, we derive the theoretical bound of maximal PCC gain from feature augmentation (e.g. when molecular and clinical predictors are combined in classifier development). Our methods are motivated and illustrated by a study using proteomics markers to classify post-kidney transplantation patients into stable and rejecting classes.

Keywords: Design; Higher criticism threshold; Large p small n; Linear discrimination; Sample size.

1. Introduction

In recent years, high-dimensional classification analysis has received heightened attention due to its importance for personalized medicine: if validated classifiers (e.g. diagnostic tests) are available, clinicians can use them to design effective treatment plans for individual patients (Hamburg and Collins, 2010). Several approaches to deriving classifiers based on high-dimensional biomarkers have been developed in the literature, and when applied to real world experiments, some promising results have been reported, e.g. Clarke *and others* (2008), Simon (2008), and Wang *and others* (2008). However, rapid technological advances enabling the collection of hundreds to millions of biomarkers from a single patient give rise to study design challenges (Mardis, 2008; Schuster, 2008), including how to determine adequate sample size to train classifiers.

*To whom correspondence should be addressed.

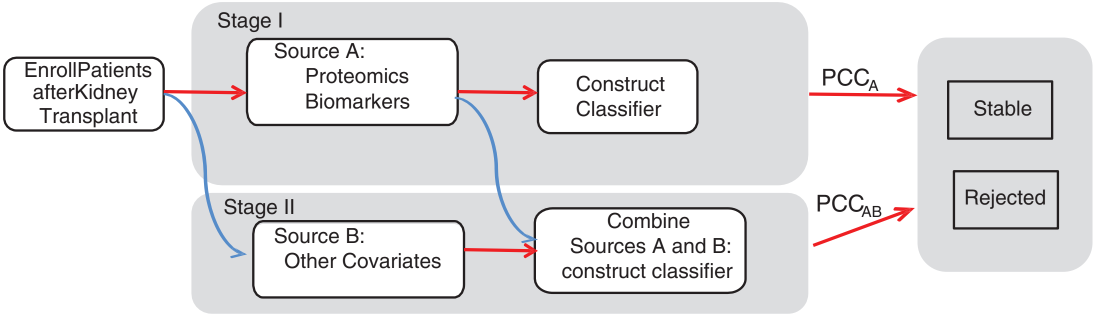

Stage I
EnrollPatients afterKidney Transplant
Source A: Proteomics Biomarkers
Construct Classifier
PCC_A
Stable
Rejected
Stage II
Source B: Other Covariates
Combine Sources A and B: construct classifier
PCC_{AB}

Fig. 1. Study flowchart for constructing classifiers of graft survival after kidney transplant. In Step I, investigators will collect $p = 108$ proteomics biomarkers using microarrays for each patient in the stable and rejecting groups, and a classifier of graft survival status will be developed. In Stage II, investigators will consider adding other clinical characteristics and patient demographics to hopefully improve classification precision.

We address two key study design issues for classification studies with high-dimensional predictors (i.e. "$n \ll p$" scenarios); namely, how to: (i) determine sample size that accounts for actual data analyses plans in advance and (ii) assess gain in classification precision associated with feature augmentation. Given space constraints, we focus only on study design related issues, an area that has received less attention. The current design literature in this area focuses on classifiers that are constructed by first screening biomarkers that may be differentially expressed across disease groups (i.e. assuming biomarkers important for classification are sparse), and subsequently combine the selected biomarkers into a classification rule. These types of classifiers rely on threshold cutoffs for selecting important features, the estimation of which needs to be accounted for in the design stage.

This work is motivated by two collaborative projects. The first is our work with the Nephrotic Syndrome Study Network (NEPTUNE), which studies molecular mechanisms for rare renal diseases. One of NEPTUNE's goals is to identify tissue-based mRNA biomarkers to classify patients into risk groups and predict disease remission. A comprehensive generalization of the NEPTUNE study design method (Gadegbeku *and others*, 2013) is frequently needed in practice.

A second collaboration is joint work with a clinician at The University of Michigan Kidney Transplantation Center, who aimed to predict patient's graft survival status (stable vs. rejecting), a measure of treatment effectiveness, after kidney transplant. The proposed study will proceed in two stages (Figure 1). First, the investigator would like to know how many transplant patients are sufficient to derive and validate a powerful classifier based on protein biomarkers. Second, the investigator would like to know if the classification prediction can be improved by adding clinical predictors such as routine measures of patient's laboratory tests (e.g. albium and hemoglobin) and demographic characteristics—i.e. the gain in prediction accuracy due to feature augmentation.

Aside from sample size determination methods that optimize hypothesis testing criteria in high-dimensional data settings (e.g. Hwang *and others*, 2002), few sample size methods for building classifiers are available. One ground breaking method for classification analysis was proposed by Dobbin and Simon (2007) (hereafter DS2007), which is based on optimizing the probability of correct classification (PCC, Mukherjee *and others*, 2003). The classifier's PCC (or sensitivity or specificity) is a more appropriate target for sample size determination for classification studies, rather than the classical concepts of Type I and Type II errors for testing differences across groups. One limitation of DS2007's method is that the threshold for feature selection is optimized for the given design parameters (e.g. number of important features and their effect size), and this threshold is treated as known in the sample size calculation. As a result, this design approach does not have a counterpart in the data analysis stage, because during analyses the true differences between groups, and thus the threshold, are unknown. Liu *and others* (2012) develop sample size determination methods for classifiers based on single nucleotide polymorphisms, which was extended by Liu *and others* (2014) to multi-class classifiers. de Valpine *and others* (2009) develop a simulation-approximation approach to determine sample size.

The benefits of feature augmentation in terms of the receiver operator curve have been investigated (Pepe *and others*, 2006; Cai and Cheng, 2008; Pfeiffer and Bur, 2008; Lin *and others*, 2011). However, there is no theoretical work specifically quantifying the amount of PCC gain due to feature augmentation, nor under which scenarios PCC gain can be maximized.

Section 2 describes the model formulation for the features, the PCC definition, and two thresholding techniques used to select important features with which the classification rule is constructed. One is the higher criticism threshold (HCT) proposed by Donoho and Jin (2009), which is particularly relevant when important features are rare and weak, and the other is a method based on cross validation (CV). Section 3 presents our proposed methods for sample size determination which incorporate thresholding techniques. We introduce a new simulation method to efficiently evaluate the PCC of HCT-based classifiers. In Section 4, we establish a novel inequality with both the upper and lower bounds for PCC gain due to feature augmentation. Section 5 illustrates the performance of three sample size determination strategies and their use in the second motivating example of predicting kidney graft status, followed by a discussion.

2. Model, PCC, and feature selection

In this section, we review existing work and modeling set up that serves as the context for our proposed sample size determination methods. Suppose the study population can be divided into two groups: Group $+1$ and Group -1. The design question is on how many subjects, n, a set of *training data* $D = \{(\boldsymbol{x}_i, y_i) \mid i = 1, \ldots, n\}$ will be collected to construct a classifier, where $y_i = \{+1, -1\}$ is the group label for subject i; population group prevalences are $P(y_i = +1) = p_1$ and $P(y_i = -1) = 1 - p_1$, respectively; and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T \in R^p$ is a high-dimensional vector of features for subject i (e.g. proteomics biomarkers). For brevity of exposition, in the rest of the paper we assume the sample size collected from each group is equal by design (e.g. stratified sampling is used), irrespective of the group prevalences in the population. Supplementary material available at *Biostatistics* online describes modifications needed when sample sizes are unequal for the groups.

We assume that features follow the multivariate normal distribution within each group with equal variances: $\boldsymbol{x}_i \mid y_i = +1 \sim N(+\boldsymbol{\mu}, \Sigma)$; and $\boldsymbol{x}_i \mid y_i = -1 \sim N(-\boldsymbol{\mu}, \Sigma)$, where the vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$, with elements $\mu_j \geqslant 0$, $j = 1, \ldots, p$, represents the signal strengths of the features. Setting $E(\boldsymbol{x}_i \mid y_i) = \boldsymbol{\mu} y_i$ is purely for notational convenience and is not needed in practice; this notation allows us to write the mean differences of features between groups in terms of a single vector, namely $2\boldsymbol{\mu}$. A higher value of μ_j suggests a better separation between two groups by feature j, and consequently feature j would be important for classification. Assuming the equality of variances is needed to construct a linear classification rule (Johnson and Wichern, 2002), which we assume at the design stage. Without loss of generality, we assume the diagonal elements of Σ equal 1, which enables us to refer to $2\boldsymbol{\mu}$ as the vector of effect sizes. In practice, this is achieved by dividing each feature by its pooled standard deviation calculated with the training data.

The dimension p of $\boldsymbol{\mu}$ may be very high; hence it is commonly assumed that only a small number of features, say, m ($\ll p$), have non-zero effect sizes. The m features are considered essential to construct a classifier, while the other $p - m$ features are noise ($\mu_j = 0$). For the ease of exposition, we reorder features such that the important m features are listed first, i.e. $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m, \boldsymbol{0}_{p-m}^T)^T$, where $\boldsymbol{0}_{p-m}$ is a zero vector of length $p - m$. The values of effect sizes, μ_j, $j = 1, \ldots, m$, are unknown at the design stage and m is supplied by subject-matter scientists. Assume that it is possible to specify a lower bound

μ_0 for $\mu_1, \ldots, \mu_m$ based on some prior research results or a certain scientific hypothesis, and replace $\boldsymbol{\mu}$ by $\boldsymbol{\mu}_0 = (\mu_0 \mathbf{1}_m^T, \mathbf{0}_{p-m}^T)^T$, where $\mathbf{1}_m$ is a vector of ones of length m. Then, we will simply state the effect size as $2\mu_0$. Under the linear classification rule and given the weighting scheme defined in (2.1), using a lower bound in the design will lead to a conservative estimate of PCC and thus sample size, which is acceptable in practice when no reliable pilot data are available to estimate $\mu_1, \ldots, \mu_m$ satisfactorily.

In this paper, we consider a linear classifier in the design stage. Constructing a linear classifier is equivalent to using training data D to derive a certain weighting scheme G that allocates weights $\boldsymbol{w} = (w_1, \ldots, w_p)^T = G(D)$. Let $\kappa = \frac{1}{2}\log((1-p_1)/p_1)$ and $\boldsymbol{a} \cdot \boldsymbol{b}$ denote the inner product of two vectors. The classification rule for a new subject is: if $\boldsymbol{w} \cdot \boldsymbol{x}_i \geqslant \kappa$, subject i is assigned to Group $+1$; otherwise to Group -1. In general, the weighting scheme G can assign non-zero weights $\boldsymbol{w}$ to all available features; however, this can harm PCC if most of them are not important. Instead, when $n \ll p$ using regularized feature selection allows us to include only important features in the classifier, thus enhancing the classifier's PCC. Feature selection is primarily driven by pairwise associations between features x_{ij} and group membership y_i. Let $Z = (z_1, \ldots, z_p)^T$ be the vector of test statistics derived from training data D. Then $z_j \sim N(0, 1)$ for unimportant features, and $z_j \sim N(\tau, 1)$ for $j = 1, \ldots, m$ where $\tau = \sqrt{n}\mu_0$ is the signal strength. A natural strategy for feature selection is to choose an appropriate threshold λ such that we only include features satisfying $|z_j| \geqslant \lambda$, $j = 1, \ldots, p$. Given threshold λ, we incorporate this feature selection mechanism into the definition of the weighting scheme:

$$w_j = 1, \quad \text{if} \quad z_j > \lambda; \quad w_j = -1, \quad \text{if} \quad z_j < -\lambda; \quad \text{and} \quad w_j = 0 \quad \text{otherwise.} \tag{2.1}$$

The threshold λ is determined empirically given D; Section 2.2 describes procedures to select it.

2.1 *Objective function and connection to sample size*

Following DS2007, we use PCC as the primary objective function for sample size determination. With two groups, the PCC is the weighted average of the classifier's sensitivity and specificity, with weights equal to the group prevalences. Under the assumed model and for fixed weights $\boldsymbol{w}$, it can be easily shown that $\text{PCC}(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, p_1) = p_1 \Phi((\boldsymbol{w} \cdot \boldsymbol{\mu} - \kappa)/\sqrt{\boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w}}) + (1 - p_1)\Phi((\boldsymbol{w} \cdot \boldsymbol{\mu} + \kappa)/\sqrt{\boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w}})$, where $\Phi(\cdot)$ is the standard normal CDF. The weights $\boldsymbol{w} = G(D)$, however, are random and depend on $G(\cdot)$ (and thus λ) and the sample size n of D.

To make the connection between PCC and sample size, it is useful to think of the PCC as dependent on sample size and the selected threshold λ, hence defining $\text{PCC}(\boldsymbol{w}_\lambda; \boldsymbol{\mu}, \boldsymbol{\Sigma}, p_1, n)$ as the PCC of a classifier built using training data on n subjects. Further, to define the optimal PCC, it is useful to think of the upper bound of the PCC among linear classifiers. A linear classifier can reach the upper bound if it is the oracle classifier, or if it is constructed from a study with infinite sample size (DS2007). This optimal classifier has $\text{PCC}_{\text{oracle}} = p_1 \Phi((\boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} - \kappa)/\sqrt{\boldsymbol{\mu} \Sigma^{-1} \boldsymbol{\mu}}) + (1 - p_1)\Phi((\boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} + \kappa)/\sqrt{\boldsymbol{\mu} \Sigma^{-1} \boldsymbol{\mu}})$, which simplifies to $\text{PCC}_{\text{oracle}} = \Phi(\sqrt{m}\mu_0)$ when $\boldsymbol{\Sigma} = I$, $p_1 = \frac{1}{2}$, and $\boldsymbol{\mu}$ is replaced by $\boldsymbol{\mu}_0 = (\mu_0 \mathbf{1}_m, \mathbf{0}_{p-m})$ at the design stage.

Clearly, a practically achievable PCC will be lower than the upper bound, and its exact value depends on how much relevant information can be extracted from the training data. At the design stage, a PCC target is set lower than $\text{PCC}_{\text{oracle}}$; for instance, DS2007 set $\text{PCC}_{\text{target}}$ as the smallest PCC that satisfies $\text{PCC}_{\text{target}} \geqslant \text{PCC}_{\text{oracle}} - 0.05$. The sample size requirement is then defined as the smallest n such that $\text{PCC}(\boldsymbol{w}_\lambda; \mu_0, n) \geqslant \text{PCC}_{\text{target}}$. If the inverse function of $\text{PCC}(n)$ could be analytically derived, then the sample size would be easily determined by $n \geqslant \text{PCC}^{-1}(\text{PCC}_{\text{target}})$. Since a closed form of $\text{PCC}^{-1}(\cdot)$ rarely exists, we employ numerical algorithms (e.g. the binary search algorithm) to determine sample size.

Finally, one must consider how to calculate PCC at the design stage. DS2007 consider an approximation of PCC to warrant its fast computation. First they compute an *optimal, fixed* threshold λ using information on the assumed μ_0, and show with Monte Carlo (MC) simulations that for a *fixed* λ the approximation $E_D[\mathrm{PCC}(\boldsymbol{w}_\lambda; \mu_0, n)] \approx \mathrm{PCC}(E_D[\boldsymbol{w}_\lambda]; \mu_0, n)$ is accurate. However, at the analysis stage, optimizing λ only depends only on the data D, since $\boldsymbol{\mu}$ is unknown. We describe two procedures to determine λ based only on simulated data and are thus advantageous because they have an actual parallel in the analysis stage.

2.2 *Feature selection procedures*

2.2.1 *CV threshold.* We consider the following straightforward K-fold CV thresholding method to determine λ given only training data D, denoted by $\lambda = \mathrm{CV}(D)$. Such λ is chosen to maximize the apparent PCC, $\widetilde{\mathrm{PCC}}(D, \lambda)$, which is a function of both threshold λ and training data D. The apparent PCC can be computed via the following steps:

(1) Follow the sampling strategy for D to divide it into K equal-sized subsets, $D^1, \ldots, D^K$ (e.g. divide cases and controls separately if stratified sampling is used to collect D).

(2) For each $q = 1, \ldots, K$, treat D^q, which has sample size n_q, as a CV testing set and the rest of the data D^{-q} as a CV training set; given a threshold value λ^*, which is one of many threshold values on a dense grid, use (2.1) to obtain the weighting $w(D^{-q}, \lambda^*)$ from the training set D^{-q}, where z-scores are calculated from dataset D^{-q} only.

(3) For $q = 1, \ldots, K$, calculate the apparent PCC based on testing set D^q, as $\widetilde{\mathrm{PCC}}(D^q, w(D^{-q}, \lambda^*)) = \sum_{(\boldsymbol{x}_i, y_i) \in D^q} \{I(\boldsymbol{x}_i \cdot w(D^{-q}, \lambda^*) \geqslant 0) I(y_i = 1) + I(\boldsymbol{x}_i \cdot w(D^{-q}, \lambda^*) < 0) I(y_i = -1)\}/n_q$.

(4) Calculate the overall apparent PCC: $\widetilde{\mathrm{PCC}}(D, \lambda^*) = (1/K) \sum_{q=1}^{K} \widetilde{\mathrm{PCC}}(D^q, w(D^{-q}, \lambda^*))$.

(5) Calculate the apparent PCC on a dense grid of values for λ^*, and select the optimal threshold λ that maximizes the overall apparent PCC: $\mathrm{CV}(D) = \mathrm{argmax}_{\lambda^* > 0} \widetilde{\mathrm{PCC}}(D, \lambda^*)$.

For an analysis employing the CV threshold, the expected PCC is calculated over the distribution of D, $E_D\{\mathrm{PCC}(G_{\mathrm{CV}(D)}(D); \boldsymbol{\mu}), n\}$. This procedure optimizes the threshold λ without bearing on knowledge of $\boldsymbol{\mu}_0$; when embedded in sample size calculations it accounts for uncertainty in λ. When features are independent and important features have the same effect size, CV thresholding with weights in (2.1) results in the optimal Bayes classification rule, except for uncertainty in λ, and thus the optimal sample size.

2.2.2 *Higher criticism threshold.* Proposed by Donoho and Jin (2009), HCT provides a data-driven approach to determine λ in a high-dimensional classification analysis. HCT determines a suitable threshold λ based on the distribution of p-values obtained from univariate tests for associations of individual features with the group assignment, and then the weighting scheme (2.1) can be applied. Let $\mathrm{HCT}(D)$ denote the HCT procedure when applied to training data D. The association test for feature x_{ij} with group y_i results in a two-sided p-value $\pi_j = 2\{1 - \Phi(|z_j|)\}$. For an unimportant feature, $\pi_j \sim \mathrm{Uniform}(0, 1)$; for an important feature the resulting π_j does not follow $\mathrm{Uniform}(0, 1)$ and tends to be smaller than those of the unimportant features. HCT only focuses on the smallest $\lceil p\alpha_0 \rceil$ p-values sorted in increasing order: $\pi_{(1)}, \ldots, \pi_{(\lceil p\alpha_0 \rceil)}$; a typical choice is $\alpha_0 = 10\%$. Donoho and Jin (2009) showed that the lth ordered p-value with $l = \mathrm{argmax}_{k=1,\ldots,\lceil p\alpha_0 \rceil} \sqrt{p}((k/p - \pi_{(k)})/\sqrt{k/p(1 - k/p)})$ provides an appropriate cutoff for feature selection: features whose p-values are less than $\pi_{(l)}$ are considered important for classification. The z-score threshold is thus $\mathrm{HCT}(D) = |\Phi^{-1}(\pi_{(l)}/2)|$. The resulting PCC is $E_D\{\mathrm{PCC}(G_{\mathrm{HCT}(D)}(D); \boldsymbol{\mu}, n)\}$.

As detailed by Donoho and Jin (2009), the theory of HCT brings new insights to the asymptotic properties of linear classifiers under the so-called rare-and-weak model, which is of interest in the context of high-dimensional classification because it gives a structure under which the number of important features m and signal strength τ vary with the total number of features p. This structure enables study of asymptotic classification feasibility. In this rare-and-weak model, m increases with p according to

$m = p^{1-\beta}$ (Donoho and Jin, 2009), where $\beta \in (0, 1)$ controls the sparsity. Similarly, instead of $\tau = \sqrt{n}\mu_0$, which becomes arbitrarily large with increasing sample size, in this model the signal strength follows $\tau = \sqrt{2r \log p}$; $r \in (0, 1)$ controls the signal strength, and $n \propto \{\log(p)\}^{\gamma}$ for some $\gamma > 0$. This implies that important features become rarer and their effect size becomes weaker when the total number of features p increases, which is regarded as a more realistic mechanism (NCI-NHGRI, 2007; Jin, 2009) than a mechanism where PCC always increases as p increases. It has been shown (Donoho and Jin, 2009; Jin, 2009) that as the number of features $p \to \infty$ the PCC of any linear classifier is characterized only by (β, r) through a certain function $\rho(\beta)$ given in Section B of supplementary material available at *Biostatistics* online: (i) when $r > \rho(\beta)$, the classification analysis is asymptotically feasible, in the sense that the PCC of the HCT linear classifier approaches to 1 as $p \to \infty$ and (ii) when $r < \rho(\beta)$, the classification analysis is asymptotically infeasible. The asymptotic result of feasibility is critical to guide the design of classification analysis. Verifying the inequality $r > \rho(\beta)$ can help investigators make a timely decision on the feasibility of a study at the planning stage (see Section 5 for an illustration).

3. Implementation

Given an approach to evaluate PCC, sample size can be determined by inverting the PCC function numerically. We thus focus on PCC estimation approaches that incorporate thresholding procedures so the resulting PCC would more closely reflect what can be achieved in practice. Section C of supplementary material available at *Biostatistics* online describes the evaluation of PCC for CV-based classifiers. Our primary contributions here focus on approaches needed for HCT-based classifiers.

Since HCT(D) is a data-driven thresholding procedure, MC simulation can be applied to evaluate $E_D\{\text{PCC}(G_{\text{HCT}(D)}(D); \boldsymbol{\mu}, n)\}$. However, because HCT$(D)$ depends on the training data D exclusively through the $\lceil p\alpha_0 \rceil$ smallest p-values, we propose a computationally fast MC algorithm that directly simulates the $\lceil p\alpha_0 \rceil$ smallest p-values from the distribution of ordered statistics instead of simulating D. The algorithm is takes the following steps:

(a) Simulate z-scores for m important features from $z_j \sim N(\mu_0\sqrt{n}, 1)$, $j = 1, \ldots, m$.
(b) Convert the above z-scores to two-sided p-values by $\pi_j = 2\{1 - \Phi(|z_j|)\}$, $j = 1, \ldots, m$.
(c) Simulate a random variable $u \sim \text{Beta}(\lceil p\alpha_0 \rceil, p - m + 1 - \lceil p\alpha_0 \rceil)$.
(d) Simulate variables $v_1, \ldots, v_{\lceil p\alpha_0 \rceil - 1}$ independently from Uniform$(0, u)$.
(e) Sort vector $(\pi_1, \ldots, \pi_m, v_1, \ldots, v_{\lceil p\alpha_0 \rceil - 1}, u)^T$ in an ascending order.

The $\lceil p\alpha_0 \rceil$ smallest values in $(\pi_1, \ldots, \pi_m, v_1, \ldots, v_{\lceil p\alpha_0 \rceil - 1}, u)^T$ have the same joint distribution as the $\lceil p\alpha_0 \rceil$ smallest p-values $(\pi_{(1)}, \ldots, \pi_{(\lceil p\alpha_0 \rceil)})^T$ derived from D (see supplementary material available at *Biostatistics* online, Section D for the proof). As a by-product, the above algorithm also supplies the z-scores $z_1, \ldots, z_m$ for the m important features. This proposed algorithm has a clear computational benefit: instead of generating np random variables needed for D (and calculating all p-values), only $\lceil p\alpha_0 \rceil + m$ variables are generated. The computational efficiency ratio is $np/(\lceil p\alpha_0 \rceil + m) \approx n/\alpha_0$ (since m is much smaller than p); e.g. for $n = 100$ and $\alpha_0 = 10\%$, the algorithm is $\approx$ 1000 times more efficient. Furthermore, the algorithm below used to calculate PCC does not require generating test data.

To evaluate $E_D\{\text{PCC}(G_{\text{HCT}(D)}(D); \boldsymbol{\mu}, n)\}$, we repeat for $k = 1, \ldots, N$ iterations:

(1) For given n, μ_0, and m, use Steps (a)–(e) to generate the $\lceil p\alpha_0 \rceil$ smallest p-values $(\pi_{(1)}, \ldots, \pi_{(\lceil p\alpha_0 \rceil)})^T$ and the corresponding z-scores for m important features.
(2) Determine the optimal threshold $\lambda = \text{HCT}(D)$.
(3) Use (2.1) to calculate weights $w_1, \ldots, w_m$ of important features using their z-scores and λ.

(4) Calculate $\mathrm{PCC}_{(k)} = p_1\Phi((\mu_0\sum_{j=1}^{m} w_j - \kappa)/\sqrt{\#w}) + (1-p_1)\Phi((\mu_0\sum_{j=1}^{m} w_j + \kappa)/\sqrt{\#w})$, where $\#w = \sum_{j=1}^{\lceil p\alpha_0\rceil} I[\pi_{(j)} < 2\{1-\Phi(|\lambda|)\}]$ is the number of elements in $(\pi_{(1)}, \ldots, \pi_{(\lceil p\alpha_0\rceil)})^T$ that are smaller than $2\{1-\Phi(|\lambda|)\}$.

Then, the fast MC estimate of $E_D\{\mathrm{PCC}(G_{\mathrm{HCT}(D)}(D); \boldsymbol{\mu}, n)\}$ is given by $(1/N)\sum_{k=1}^{N}\mathrm{PCC}_{(k)}$.

Correlated features. When features are correlated, estimating PCC is more challenging. One difficulty pertains to computing the denominator, $\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}$, within the PCC formula, $\mathrm{PCC}(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, p_1) = p_1\Phi((\boldsymbol{w}\cdot\boldsymbol{\mu} - k)/\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}) + (1-p_1)\Phi((\boldsymbol{w}\cdot\boldsymbol{\mu} + k)/\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}})$. DS2007 proposed replacing this quantity with an upper bound $\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} \leqslant e\boldsymbol{w}^T\boldsymbol{w}$, where e is the largest eigenvalue of $\boldsymbol{\Sigma}$. This bound could potentially be applied at Step 4 of the HCT-based algorithm above when calculating $\mathrm{PCC}_{(k)}$. However, this does not work for the HCT method since Step 1 relies on independence to prove that the u used to generate the $\lceil p\alpha_0\rceil$ smallest p-values follows a Beta distribution. Hence, we instead evaluate the expected PCC using an alternative MC simulation strategy. The simulation strategy follows Steps 1–4 as above, with the following modifications. First, we specify an assumed working correlation structure $\boldsymbol{\Sigma}$ for the features. In Step 1, we now use this assumed $\boldsymbol{\Sigma}$ to generate p correlated features on n subjects, and compute z statistics and accompanying p-values. The choice of structure for $\boldsymbol{\Sigma}$ (e.g. block diagonal) may be informed by substantive knowledge, if available, and the magnitude of the correlations should be varied to assess its impact on the sample size calculation. Steps 2 and 3 of the algorithm remain unchanged to reflect that at the analysis stage the features are screened using pairwise associations and treated as independent. In Step 4, we evaluate $\mathrm{PCC}_{(k)} = p_1\Phi((\mu_0\sum_{j=1}^{m} w_j - \kappa)/\sqrt{\#w}) + (1-p_1)\Phi((\mu_0\sum_{j=1}^{m} w_j + \kappa)/\sqrt{\#w})$, where $\#w$ is an estimate of $\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}$ based on the working correlation and is defined as $\#w = \sum_{(j,j'):w_j\neq 0, w_{j'}\neq 0}\sigma_{jj'}w_j w_{j'}$ and $\sigma_{jj'}$ denotes the (j, j') entry of $\boldsymbol{\Sigma}$. As with the eigenvalue approach, this approximation relies on the working correlation structure $\boldsymbol{\Sigma}$; however, this approach is less conservative than using the eigenvalue-based bound (see Illustrations section). The MC approach for the CV method with correlated features similarly relies on generating correlated data (see supplementary material available at *Biostatistics* online).

4. Feature augmentation

Because in practice multiple sources of features are typically collected, a key study design question is to investigate the potential PCC gain resulting from adding new sources of features to the classification analysis. For simplicity, let us focus on two sources of features (e.g. molecular biomarkers and clinical variables). Denote features already in the study by Type A and the new set by Type B, with respective dimensions p_A and p_B. For subject i, we collect measurements $\boldsymbol{x}_i^A \in R^{p_A}$ and $\boldsymbol{x}_i^B \in R^{p_B}$. As in Section 2, assume that the joint conditional distribution for features $\boldsymbol{x}_i^A$ and $\boldsymbol{x}_i^B$ is:

$$\left.\begin{pmatrix}\boldsymbol{x}_i^A\\ \boldsymbol{x}_i^B\end{pmatrix}\right| y_i \sim N\left(y_i\begin{pmatrix}\boldsymbol{\mu}^A\\ \boldsymbol{\mu}^B\end{pmatrix}, \begin{pmatrix}\Sigma^A & \Sigma^{AB}\\ \Sigma^{BA} & \Sigma^B\end{pmatrix}\right)$$

where $\boldsymbol{\mu}^A$ and $\boldsymbol{\mu}^B$ and Σ^A, Σ^B, and $\Sigma^{AB} = (\Sigma^{BA})^T$ are the respective effect size vectors, variance, and covariance matrices. Here we do not place any restrictions on either the effect size vectors (e.g. sparsity is not assumed) or the variance matrices.

To study PCC gain, we need to consider the PCC of linear classifiers in three cases, including (i) Type A features only; (ii) Type B features only; and (iii) Type A features augmented with Type B features. Denote the respective weights in Cases (i) and (ii) by w_A and w_B; we do not place any assumptions on these weights, e.g. they may be derived from any thresholding procedure for feature selection. Conditioning on the weights w_A and w_B, and assuming group prevalence is $p_1 = 1/2$, the PCC of the classifier in Case (i) is $\mathrm{PCC}_A = \Phi((\boldsymbol{\mu}^A\cdot\boldsymbol{w}_A)/\sqrt{\boldsymbol{w}_A^T\Sigma^A\boldsymbol{w}_A})$; in Case (ii) is $\mathrm{PCC}_B = \Phi((\boldsymbol{\mu}^B\cdot\boldsymbol{w}_B)/\sqrt{\boldsymbol{w}_B^t\Sigma^B\boldsymbol{w}_B})$; and finally, in Case (iii) is $\mathrm{PCC}_{AB} = \Phi((\boldsymbol{\mu}^A \cdot \boldsymbol{w}_A + \boldsymbol{\mu}^B \cdot \boldsymbol{w}_B)/\sqrt{\boldsymbol{w}_A^T \Sigma^A \boldsymbol{w}_A + \boldsymbol{w}_B^T \Sigma^B \boldsymbol{w}_B + \boldsymbol{w}_A^T \Sigma^{AB} \boldsymbol{w}_B})$. When $\Sigma^{BA} = 0$, the term $\boldsymbol{w}_A^T \Sigma^{AB} \boldsymbol{w}_B$ drops from the denominator and we obtain $\mathrm{PCC}_{AB,\mathrm{IND}}$. In Section E of supplementary material available at *Biostatistics* online, we prove that:

$$\min(\mathrm{PCC}_A, \mathrm{PCC}_B) \leqslant \mathrm{PCC}_{AB,\mathrm{IND}} \leqslant \Phi(\sqrt{2} \cdot \Phi^{-1}(\max(\mathrm{PCC}_A, \mathrm{PCC}_B))). \tag{4.1}$$

The first equality holds when the relative variance of the linear predictors goes to 0, i.e. $\mathrm{Var}(\boldsymbol{w}_k \cdot \boldsymbol{x}_i^k)/\mathrm{Var}(\boldsymbol{w}_j \cdot \boldsymbol{x}_i^j) = \boldsymbol{w}_k^T \Sigma^k \boldsymbol{w}_k / \boldsymbol{w}_j^T \Sigma^j \boldsymbol{w}_j \to 0$ where $(j, k) = (A, B)$ if $\mathrm{PCC}_A < \mathrm{PCC}_\mathrm{B}$ and $(j, k) = (B, A)$ if $\mathrm{PCC}_A \geqslant \mathrm{PCC}_B$. In the latter case, the equality is reached when $\mathrm{PCC}_A = \mathrm{PCC}_B$ and the linear predictors have equal variance $w_A^T \Sigma^A w_A = w_B^T \Sigma^B w_B$. Inequality (4.1) provides the upper bound of the PCC of the classifier when the linear predictors are combined into a new classification rule. If either PCC_A or PCC_B approach 1, the upper bound will approach 1.

In practice, the features may be correlated, $\Sigma^{BA} \neq 0$, thus the linear predictors $\boldsymbol{x}^A \cdot \boldsymbol{w}_A$ and $\boldsymbol{w}_B \cdot \boldsymbol{x}_i^B$ will too. Given the monotonicity of $\Phi(\cdot)$ and that the covariance of the linear predictors, $\boldsymbol{w}_A^T \Sigma^{AB} \boldsymbol{w}_B$, appears in the denominator PCC_{AB}, the PCC gain will depend on the sign of the correlation: $\mathrm{PCC}_{AB,+} \leqslant \mathrm{PCC}_{AB,\mathrm{IND}} \leqslant \mathrm{PCC}_{AB,-}$. In Section E of supplementary material available at *Biostatistics* online, we also prove that $\min(\mathrm{PCC}_A, \mathrm{PCC}_B) \leqslant \mathrm{PCC}_{AB,+}$, and that $\mathrm{PCC}_{AB,-} \leqslant \Phi((|\delta| + 1)/|1 - \delta| \cdot \Phi^{-1}(\max(\mathrm{PCC}_A, \mathrm{PCC}_B)))$, where $\delta = \sqrt{\boldsymbol{w}_A^T \Sigma^A \boldsymbol{w}_A / \boldsymbol{w}_B^T \Sigma^B \boldsymbol{w}_B}$ is the relative standard deviation of the linear predictors. Finally, in Section E of supplementary material available at *Biostatistics* online we also prove inequality (4.1) for any proportion $p_1 \in (0, 1)$ when optimal weights are used in classifier construction.

5. Illustrations

PCC estimation and sample size determination given effect size and number of important features. For a given effect size, the PCC evaluated at the design stage will depend on a pre-specified thresholding procedure and in turn impact the sample size. Thus, we first illustrate the estimated PCC using the DS2007 method, and using the CV and HCT thresholding methods. Figure 2 shows the PCC as a function of sample size when the number of available features is $p = 500$ or $p = 10\,000$. The PCC estimated by the DS method is always the highest, primarily because it uses the true effect size to choose the optimal threshold λ. However, due to its reliance on $\boldsymbol{\mu}_0$ to obtain the optimal threshold at the design stage, the DS method has no counterpart in actual data analysis. On the other hand, the PCC estimated by the CV and the HCT methods rely only on the simulated data to estimate the threshold, which introduces uncertainty in the feature selection threshold, and thus yields lower PCC estimates. When the classification problem is more difficult (e.g. $m = 1$), HCT yields higher PCC than CV as expected (Donoho and Jin, 2008). Since PCC estimates from CV and HCT reflect more closely the achievable performance of the corresponding classifiers in real applications, the sample size estimates from these methods would better approximate the sample size required in practice.

Figure 3 shows the sample size requirements for a range of effect sizes, $m = 1$ or 10, and $\mathrm{PCC}_{\mathrm{target}} = \mathrm{PCC}_{\mathrm{oracle}} - \gamma$ with $\gamma = 0.05$. In general, sample sizes obtained from the DS method are consistently lower than with CV or HCT methods, as can be expected given in Figure 2. Sample sizes become comparable (difference $\leqslant$2) when the effect size is large. However, when the features are relatively weaker, then the DS method will tend to underestimate the needed sample size. Figure 2 also explains the facts that, for a fixed $\mathrm{PCC}_{\mathrm{target}}$, the sample sizes from HCT shown in Figure 3 are lower for scenarios when features are rarer (i.e. $m = 1$) and lower for CV when features are less rare ($m = 10$). Hence, only the proposed HCT approach will give sufficient sample size for cases when features are relatively weaker and rarer, without being overly conservative (CV method is conservative in those cases, since the HCT classifier can achieve

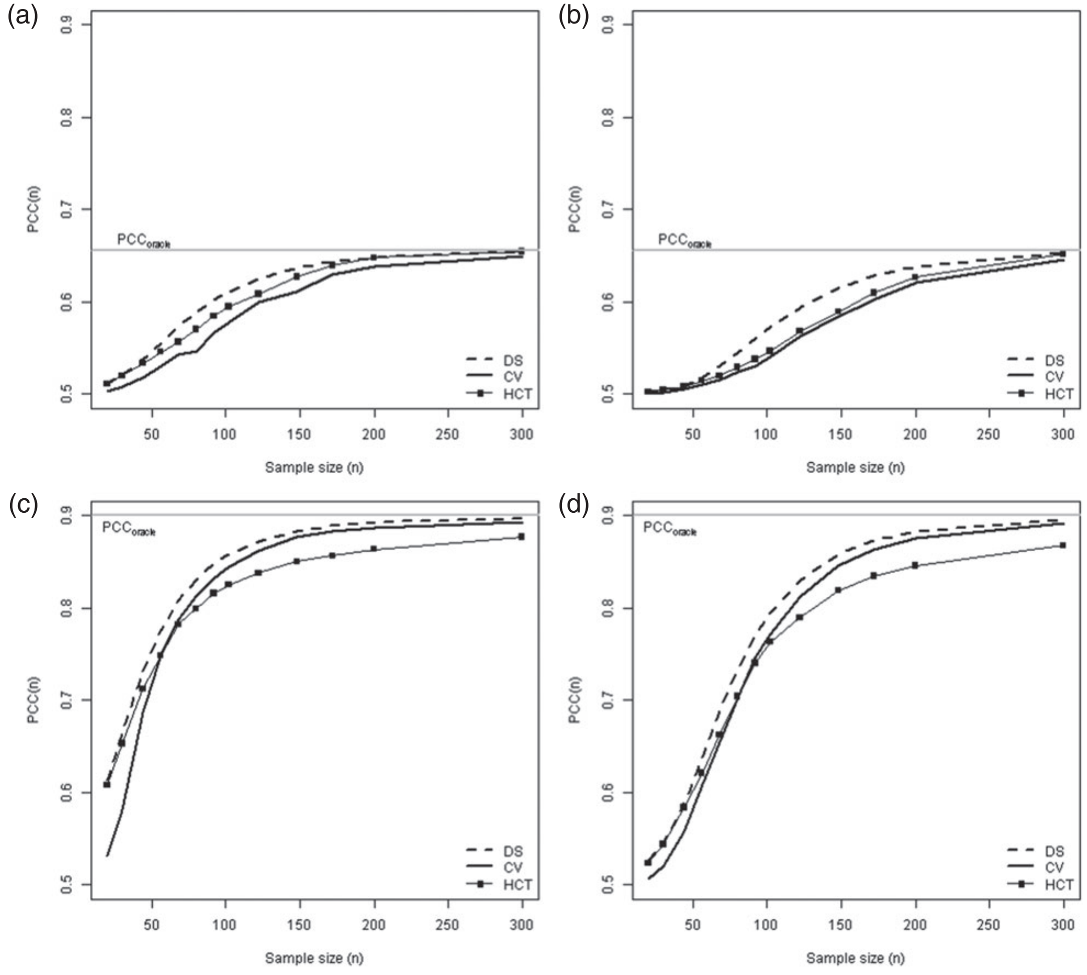


Fig. 2. PCC estimates as a function of sample size estimated by DS, CV, and HCT methods, assuming the minimal effect size of important features is $2\mu_0 = 0.8$ and group prevalences are $P(Y = +1) = P(Y = -1) = 0.5$. (a) $p = 500$, $m = 1$, (b) $p = 10\,000$, $m = 1$, (c) $p = 500$, $m = 10$, and (d) $p = 10\,000$, $m = 10$. In (a) and (b), important features are rarer ($m = 1$ important feature) compared with (c) and (d) ($m = 10$), which results in a marked difference in PCC$_{\text{oracle}}$ (gray horizontal line). The PCC estimated using the DS method is always higher for a given sample size, leading to lower sample size estimates. When features are rarer, (a) and (b), HCT gives a higher PCC, leading to lower sample size requirements. When features are less rare ($m = 10$), selecting features using HCT (CV) leads to lower sample size requirements for lower (higher) PCC targets compared with CV (HCT), given the crossing of the PCC curves. PCC estimates for CV and HCT are obtained using MC simulations using the algorithms described in Section 3, with 500 replicates for CV and 1000 replicates for HCT.

the target PCC with a lower sample size). These assertions are verified with MC simulations shown in Table 1 (top rows where $\boldsymbol{\Sigma} = I$).

It is evident from Figures 2 and 3, and Table 1 that it is important to not only calculate the sample size based on the PCC estimates that are achievable by statistical methods at the analysis stage, but also to select an analysis approach that can more efficiently attain the target PCC under a given parameter space. In rare-and-weak cases, for example, the HCT-based classifier has been shown to perform better (Donoho and Jin,

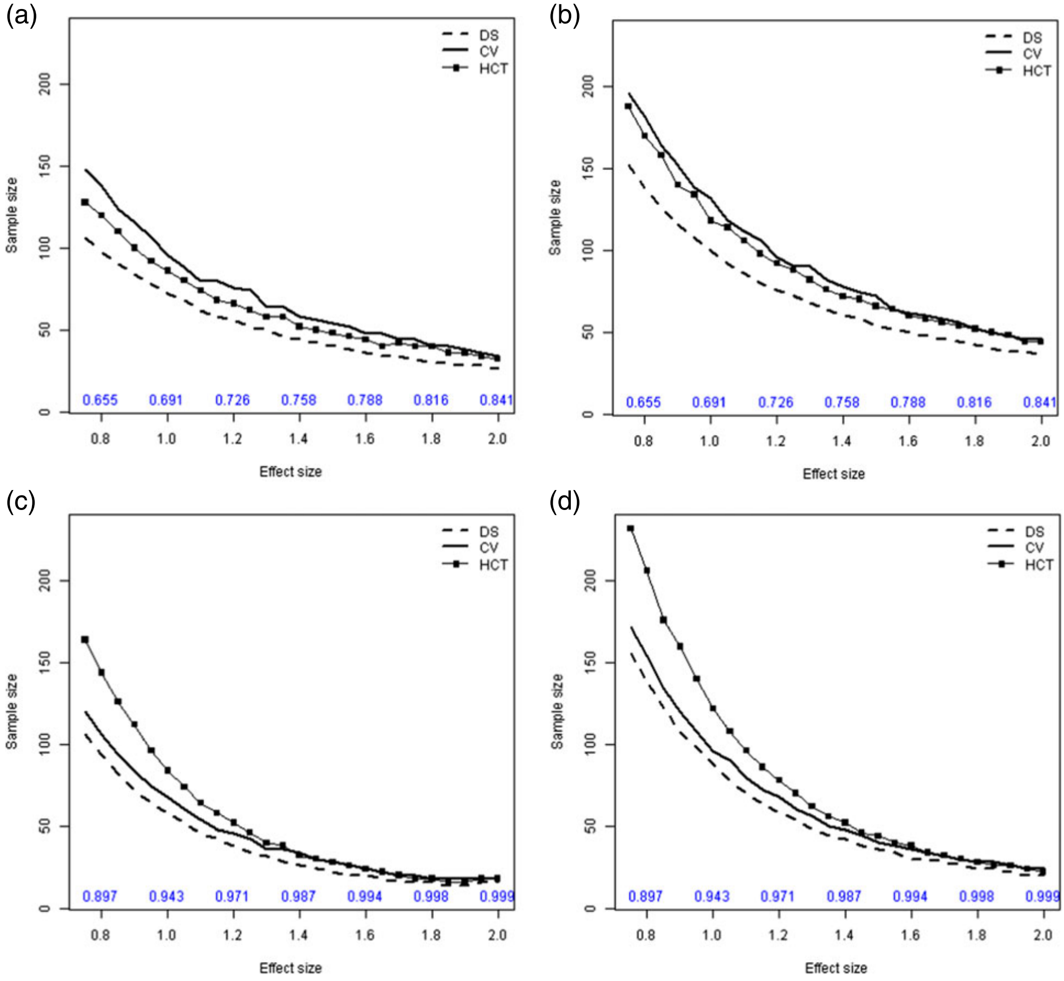


Fig. 3. Sample size requirements estimated using DS, CV, and HCT design methods for a range of effect sizes ($=2\mu_0$). (a) $p=500$, $m=1$, (b) $p=10\,000$, $m=1$, (c) $p=500$, $m=10$, and (d) $p=10\,000$, $m=10$. For each effect size and combination of m and p, the $\text{PCC}_{\text{oracle}}$ is shown as the inset value on the x-axis, and the target PCC is set as $\text{PCC}_{\text{target}} = \text{PCC}_{\text{oracle}} - 0.05$. Sample size estimates for CV and HCT are obtained by numerically inverting the PCC function and selecting the smallest n that satisfies $\text{PCC}(n) \geqslant \text{PCC}_{\text{target}}$; $\text{PCC}(n)$ is estimated using the MC algorithms described in Section 3 with 500 replicates for CV and 1000 for HCT. The sample size required decreases as effect sizes of important features increase, even in high target PCC cases. Sample sizes obtained with the DS method are lower, but, as shown in Table 1, underestimate the required sample size particularly for rare-and-weak features.

2008), and thus we recommend determining sample sizes using our proposed HCT sample size calculator in these scenarios.

Sample size calculations when features are correlated. The bottom part of Table 1 gives the sample sizes computed from each method under different structures for $\mathbf{\Sigma}$ ($\mathbf{\Sigma} \neq I$). The DS2007 method using the correction based on the largest eigenvalue sometimes yields prohibitive sample sizes ($n > 1000$, denoted as NA), because of the excessively large maximum eigenvalue of $\mathbf{\Sigma}$. Both CV- and HCT-based methods give sample sizes at which the target PCC is achieved. It is worth noting that when important features are

Table 1. *Sample size n calculated by DS, CV, and HCT design methods using the specified* $\mathrm{PCC}_{\mathrm{target}}$, *and differences between the target and what can be achieved in practice*, $\Delta\mathrm{PCC} = \mathrm{PCC}_{\mathrm{target}} - \mathrm{PCC}_{\mathrm{achieved}}$

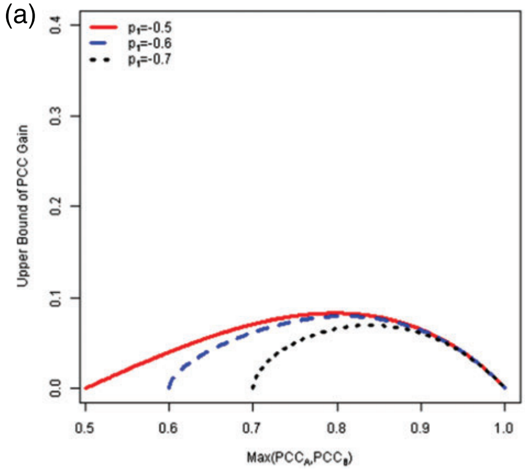
					n	ΔPCC	ΔPCC	n	ΔPCC	ΔPCC	n	ΔPCC	ΔPCC
$\mathbf{\Sigma}$	m	p	$2\mu_0$	$\mathrm{PCC}_{\mathrm{target}}$	DS	CV	HCT	CV	CV	HCT	HCT	CV	HCT
I_{500}	1	500	0.8	0.605	98	0.035	0.018	134	−0.001	−0.014	118	0.019	−0.007
			1.2	0.676	56	0.045	0.023	76	−0.003	−0.02	68	0.020	−0.002
			1.6	0.738	36	0.047	0.028	50	−0.011	−0.023	46	0.008	−0.008
$I_{10\,000}$	1	10 000	0.8	0.605	138	0.033	0.026	182	−0.001	−0.013	170	0.011	−0.001
			1.2	0.676	76	0.034	0.032	96	−0.005	−0.005	94	0.005	0.000
			1.6	0.738	50	0.043	0.042	62	0	−0.005	62	0.000	−0.005
I_{500}	10	500	0.8	0.847	94	0.014	0.031	102	0	0.025	146	−0.029	−0.002
			1.2	0.921	38	0.015	0.027	44	−0.003	0.011	52	−0.015	−0.005
			1.6	0.944	20	0.025	0.005	24	−0.004	−0.01	24	−0.004	−0.010
$I_{10\,000}$	10	10 000	0.8	0.847	138	0.01	0.032	152	0.001	0.021	202	−0.034	−0.001
			1.2	0.921	58	0.021	0.032	66	−0.006	0.014	76	−0.019	0.000
			1.6	0.944	30	0.031	0.027	36	−0.008	−0.001	36	−0.008	−0.001
Σ_2	1	500	1.6	0.738	NA	–	–	44	−0.000	0.041	62	−0.034	−0.006
Σ_3				0.738	NA	–	–	46	−0.005	0.060	168	−0.048	0.004
Σ_4				0.738	NA	–	–	46	−0.002	0.054	160	−0.044	0.004
Σ_1	10	500	1.6	0.762	26	−0.002	−0.037	26	−0.002	−0.037	18	0.064	−0.013
Σ_2				0.762	26	0.029	−0.015	30	0.000	−0.030	24	0.032	−0.010
Σ_3				0.762	NA	–	–	28	0.001	0.028	44	−0.039	−0.010
Σ_4				0.762	NA	–	–	26	0.002	0.026	42	0.044	−0.006

The scenarios and respective sample sizes shown here are a subset of those shown in Figure 3 (see Figure 3 legend for details on how sample sizes are obtained). Given the computed sample size n, ΔPCC was computed by generating 1000 training datasets of size n and test datasets of size 100; training and test datasets were generated according to the model defined by $\mathbf{\Sigma}$, m, p, and $2\mu_0$. The average of the ΔPCC across the 1000 replicates are shown. The $\mathbf{\Sigma}$ are block diagonal as follows: I_p is a $p \times p$ identity; Σ_1 has first block being 10×10 compound symmetry structure and correlation 0.80, denoted by $CS_{10}(0.8)$, and second block I_{490}; Σ_2 has 50 blocks of $CS_{10}(0.8)$; Σ_3 has first block $CS_{10}(0.8)$, second block $CS_{240}(0.8)$, and third block I_{250}; Σ_4 has the first block $CS_{250}(0.8)$, and the second block I_{250}. NA indicates sample size estimates from DS2007 eigenvalue method were prohibitive, $n > 1000$.

very rare (e.g. $m = 1$), the HCT-based method yields conservative sample sizes whereas the CV method can achieve the same target PCC with lower sample sizes.

Feature augmentation. Figure 4 illustrates the upper bound (left panel) of the PCC and PCC gain (right column) due to feature augmentation discussed in Section 4. First, in the case when features are independent (Figure 4(a)), we note that if both PCC_A and PCC_B are small (or one is large), then $\max(\mathrm{PCC}_A, \mathrm{PCC}_B)$ will be small (or large). Hence, the upper bound of the PCC of classifiers with both Type A and Type B features is only slightly higher than $\max(\mathrm{PCC}_A, \mathrm{PCC}_B)$. Combining two sets of features where both are very good or both are poor does not greatly improve PCC. If both types of features are of medium quality (e.g. both PCC_A and PCC_B are in the medium range around 0.8), then we could obtain the highest gain (at most 10%) in PCC by feature augmentation. When features are negatively correlated, the PCC gain can be substantial (Figure 4(b)).

An application. We demonstrate the proposed methods using the kidney transplant study (Figure 1). In Stage I, the investigator hypothesizes that among $p = 108$ proteins, at least $m = 10$ of them are likely informative for predicting graft survival status. Pilot data, $n = 20$, showed an effect size of approximately 0.8 (i.e. $\mu_0 \approx 0.4$). Given these design parameters, the signal strength is $\tau \approx 0.4\sqrt{20}$; the sparsity parameter is $\beta = 1 - \log m / \log p = 0.51$; and the strength parameter r lies above the feasibility boundary: $r \approx \tau^2 / 2 \log p = 0.1903 > 0.01 = \rho(\beta)$. Hence, the classification problem feasible (see Section B



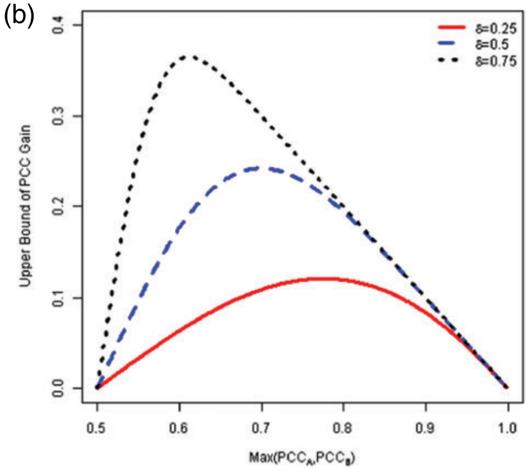


Fig. 4. The upper bounds of PCC gain $=$ PCC$_{AB}$ $-$ max(PCC$_A$, PCC$_B$) when two sets of features, A and B, are combined compared with using a single type of features. In (a), features are independent, $\mathbf{\Sigma}_{AB} = 0$ (all p_1), or $p_1 = 0.5$ and have $w_A^T \mathbf{\Sigma}_{AB} w_B \geqslant 0$. The upper bound is attained when the linear predictors have equal variance. In (b), $\mathbf{\Sigma}_{AB} \neq 0$, group 1 prevalence is $p_1 = 0.5$, and δ gives the relative standard deviation of the linear predictors. The upper bound shown in (b) is attained when the linear predictor constructed from features A and B are perfectly negatively correlated.

of supplementary material available at *Biostatistics* online), and we can proceed to calculate sample size requirements for given PCC targets (Figure 5(a)).

For Stage II, the investigator considers improving the PCC of the classifier with proteomics biomarkers only say, PCC$_A$ $= 0.7$, by incorporating an additional set of features, including proteinuria, GFR, hematuria, albium, and cholesterol. Figure 5(b) shows a region describing the achievable PCC with both types of features (PCC$_{AB,\mathrm{IND}}$ for various values of the PCC with the additional features alone, i.e. PCC$_B$). Substantial enhancements to PCC occur when the second set of features is at least as informative as the proteomics biomarkers (PCC$_B$ $\geqslant$ PCC$_A$).

6. DISCUSSION

We addressed two study design questions for studies using high-dimensional features for classification. First, we developed sample size determination strategies for CV- and HCT-based classifiers. Our strategies incorporate uncertainty of feature selection thresholds within the PCC calculation, which is particularly relevant when important features are hypothesized to be rare and weak. We proposed a computationally efficient algorithm based on order statistics to compute the PCC, and thus the sample size requirements, for the HCT-based classifier. Second, we established an inequality for the upper and lower bounds of the achievable PCC associated with feature augmentation. The approaches were illustrated with numerical examples and a practical study, and are implemented in our `R` package `HDDesign` (available at https://cran.r-project.org/).

Our proposed methods can be improved in the following directions. Classification of more than two groups commonly appears in clinical studies, thus extensions in this direction are of great importance. Strong deviations from linearity (e.g. U-shaped associations) may undermine the applicability of the proposed approaches. In this case, it may be possible to categorize the predictors and apply and/or extend the study design methods of Liu *and others* (2012) to the case of rare-and-weak features. It is also of interest to further investigate how correlations among features may be effectively incorporated into the sample size

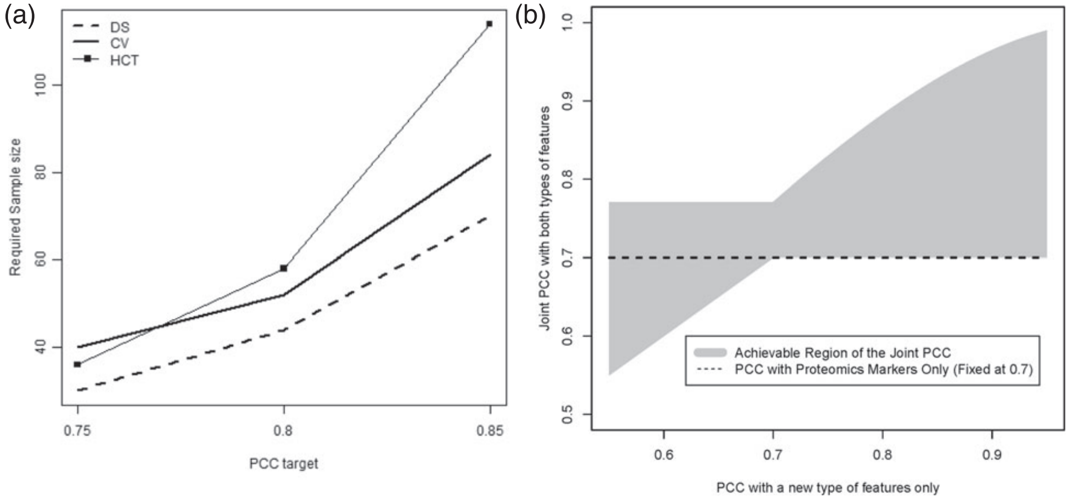


Fig. 5. Application: study design for predicting graft survival after kidney transplant. (a) As expected, a larger sample size is required when a higher PCC target is chosen and other design parameters are held constant ($\text{PCC}_{\text{oracle}} = 0.90$). The DS method yields the lowest sample size requirements, but these may underestimate the needed sample size (see Table 1); whether HCT or CV require larger sample size depends on the target PCC. (b) PCC with the proteomics markers (PCC_A) is fixed at 0.7 (dashed line). If new features are not as informative as the proteomics markers ($\text{PCC}_B \leqslant \text{PCC}_A$), combining both sets of features leads to a limited improvement of the classifier, and in some cases data augmentation might actually degrade the classifier (shadowed area below 0.7) due to the noise introduced by low-quality features in the new data source (e.g. some proteins can be measured with substantial errors if urine samples are not stored under stringent conditions). If the new features are more informative ($\text{PCC}_B \geqslant \text{PCC}_A$), incorporating them can substantially enhance the PCC.

determination. We proposed to directly plug in an assumed working correlation matrix within the CV- and HCT-based approaches. As expected, positive correlations among features result in larger required sample sizes, although our not as large as DS2007's preliminary eigenvalue-based approach. Nevertheless, our approach requires specifying sensible working correlation structures at the design stage, which may be difficult to obtain in practice. Varying the structure and magnitude of the correlations based on available scientific knowledge is needed with our proposed approach. Further improvements in this direction may be possible by using the innovated HCT suggested by Hall and Jin (2010), or by developing sample size determination methods based on regularized regression-based approaches that do not require pre-filtering and hence do not rely on the marginal effects of the features. However, developing sample size calculations using regression-based procedures (e.g. LASSO) would require specifying the adjusted effect sizes and, importantly, quantifying the uncertainty in feature selection, which remains an open problem in high-dimensional inference.

In summary, we advocate the use of sample size determination methods that match, as closely as possible, the analytic approaches that will be actually applied at the data analysis stage and that capitalize on prior knowledge of the underlying mechanism of interest. If the important features have strong signals, both HCT- and DS-based approaches provide adequate sample size calculations, and there is little difference between them. Given that the HCT method is computationally fast and accounts for uncertainty in the feature selection threshold, it is recommended in practice. If the important features are relatively abundant but weak, we recommend the CV approach as it gives the least conservative sample size, albeit computationally intensive. If the important features are rare and weak, we recommend the HCT-based approach since it provides desired sample sizes with little conservatism, and is computationally efficient.

In summary, our work builds upon and further advances the pioneering work of DS2007, for sample, size determination in high-dimensional classification problems.

Supplementary material

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The publication's contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or US EPA. *Conflict of Interest:* None declared.

Funding

M.W. and B.N.S. acknowledge NIH grant R21DA024273 for salary support during the initial conduct of this study. P.X.K.S.'s research is funded in part by NIH U54-DK-083912-05 and NSF DMS-1513595. This work was also partially funded by grants NIH/EPA P20 ES018171/RD83480001 and P01 ES022844/RD83543601.

References

Cai, T. and Cheng, S. (2008). Robust combination of multiple diagnostic tests for classifying censored event times. *Biostatistics* **9**(2), 216–233.

Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A. and Wang, Y. (2008). The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews. Cancer* **8**(1), 37–49.

de Valpine, P., Bitter, H. M., Brown, M. P. S. and Heller, J. (2009). A simulation-approximation approach to sample size planning for high-dimensional classification studies. *Biostatistics* **10**(3), 424–435.

Dobbin, K. K. and Simon, R. M. (2007). Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics* **8**(1), 101–117.

Donoho, D. and Jin, J. (2008). Higher criticism thresholding: optimal feature selection when useful features are rare and weak. *PNAS* **105**(39), 14790–14795.

Donoho, D. and Jin, J. (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* **367**(1906), 4449–4470.

Gadegbeku, C. A., Gipson, D. S., Holzman, L., Ojo, A. O., Song, P. X., Barisoni, L., Sampson, M. G., Kopp, J. B., Lemley, K. V., Nelson, P. J. *and others*. (2013). Design of the nephrotic syndrome study network (neptune): a multi-disciplinary approach to understanding primary glomerular nephropathy. *Kidney International* **83**(4), 749–756.

Hamburg, M. A. and Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine* **363**(4), 301–304.

Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics* **38**(3), 1686–1732.

Hwang, D., Schmitt, W. A., Stephanopoulos, G. and Stephanopoulos, G. (2002). Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics (Oxford, England)* **18**(9), 1184–1193.

Jin, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proceedings of the National Academy of Sciences* **106**(22), 8859–8864.

Johnson, R. A. and Wichern, D. W. (2002) *Applied Multivariate Statistics*, 5th edition. Upper Saddle River, NJ: Prentice-Hall, Inc.

Lin, H., Zhou, L., Peng, H. and Zhou, X. H. (2011). Selection and combination of biomarkers using ROC method for disease classification and prediction. *Canadian Journal of Statistics* **39**(2), 324–343.

Liu, X., Wang, Y., Rekaya, R. and Sriram, T. N. (2012). Sample size determination for classifiers based on single-nucleotide polymorphisms. *Biostatistics* **13**(2), 217–227.

Liu, X., Wang, Y. and Sriram, T. N. (2014). Determination of sample size for a multi-class classifier based on single-nucleotide polymorphisms: a volume under the surface approach. *Journal of Biomedical Informatics* **15**, 190.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24**(3), 133–141.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. R. and Mesirov, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* **10**(2), 119–142.

NCI-NHGRI, Working group on replication in association studies (2007). Replicating genotype-phenotype associations. *Nature* **447**(7145), 655–660.

Pepe, M. S., Cai, T. and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* **62**(1), 221–229.

Pfeiffer, R. M. and Bur, E. (2008). A model free approach to combining biomarkers. *Biometrical Journal* **50**(4), 558–570.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods* **5**(1), 16–18.

Simon, R. (2008). Development and validation of biomarker classifiers for treatment selection. *Journal of Statistical Planning and Inference* **138**(2), 308–320.

Wang, Y., Miller, D. J. and Clarke, R. (2008). Approaches to working in high-dimensional data spaces: gene expression microarrays. *British Journal of Cancer* **98**(6), 1023–1028.

[*Received December 11, 2014; revised April 2, 2016; accepted for publication April 3, 2016*]