1-7-2005

Combining Predictors for Classification using the Area Under the ROC Curve

Margaret S. Pepe
University of Washington, mspepe@u.washington.edu

Tianxi Cai
Harvard University, tcai@hsph.harvard.edu

Zheng Zhang
Emory University, zzhang7@sph.emory.edu

Gary M. Longton
Fred Hutchinson Cancer Research Center, glongton@fhcrc.org

1. Introduction

The use of clinical and laboratory data to detect conditions and predict patient outcomes is a mainstay of medical practice. Classification and prediction are equally important in other fields of course (e.g., meteorology, economics, computer science) and have been subjects of statistical research for a long time. The field is currently receiving more attention in medicine, in part because of biotechnologic advancements that promise accurate non-invasive modes of testing. Technologies include gene expression arrays, protein mass spectrometry and new imaging modalities. These can be used for purposes such as detecting subclinical disease, evaluating the prognosis of patients with disease and predicting their responses to different choices of therapy. Statistical methods have been developed for assessing the accuracy of classifiers in medicine (Zhou, Obuchowski, and McClish 2002; Pepe 2003), although this is an area of statistics that is evolving rapidly.

In practice there may be multiple sources of information available to assist in prediction. For example, clinical signs and symptoms of disease may be supplemented by results of laboratory tests. As another example, it is expected that multiple biomarkers will be needed for detecting subclinical cancer with adequate sensitivity and specificity (Pepe et al. 2001). The starting point for the research we describe in this paper is the need to combine multiple predictors together, somehow, in order to predict a binary outcome.

Case-control study designs are frequently employed. Study subjects are selected on the basis of the binary outcome, D. Then collect data are collected on their P predictor variables, $Y = \{Y_1, \ldots, Y_P\}$. Such retrospective designs usually require far smaller sample sizes than prospective studies, at least when the outcome is rare (or very prevalent). In medicine, case-control studies are often employed for classifier development while much larger prospective studies are undertaken only at the final phases of evaluating the classifier (Pepe 2003, Chapter 8; Pepe et al. 2001). Therefore, we focus on statistical methods for combining predictors

that can accommodate case-control designs.

To gauge the performance of a classifier we employ the traditional measures of classification accuracy that are used in medicine, namely the true- and false-positive rates (TPR and FPR). Also known as sensitivity and 1-specificity, respectively, TPR and FPR are defined as

$$\text{TPR} = \text{P[classifier positive} \mid \text{outcome positive]}$$
$$\text{FPR} = \text{P[classifier positive} \mid \text{outcome negative]}.$$

The importance of reporting both dimensions is widely recognized since the consequences of false-negative and false-positive errors are often very different and hard to quantify. One dimensional summary measures such as the overall misclassification rate or odds ratio are rarely used in practice and can in fact provide misleading results (Pepe et al. 2004). The positive and negative predictive value is another popular two-dimensional measure of accuracy. However, since it cannot be assessed directly from case-control studies, we do not consider it further here.

When multiple predictors are available or predicotrs are non-binary, it turns out that it is enough to consider classification rules based on a scalar valued function of the predictors $L(Y)$. Not only is this a large intuitively appealing class, but we note in Section 2 that it includes the optimal rules, namely those that are defined by the risk score function, $P[D = 1|Y]$, or a monotone function of it. The receiver operating characteristic (ROC) curve generalizes the notions of (TPR, FPR) from binary classifiers to scalar valued classifiers. It is a plot of TPR(c) versus FPR(c) for the rule using c as a threshold for defining a positive classification, $L(Y) > c$, $c \in (-\infty, \infty)$. The ROC curve has become the standard description of classification accuracy for scalar valued classifiers, like biomarkers (Baker

2003). Amongst its appealing properties is the fact that it provides an appropriate common scale for comparing predictors (or scalar valued combinations of them) even if the predictors themselves are not measured in the same measurement units. See Pepe (2003 chapters 4, 5 and 6) for a review of ROC methodology.

In summary, this paper addresses the question of how to combine multiple predictors into a score, i.e., a scalar valued function, when the goal is to use that score for classification. The method used for evaluating classification accuracy of the score is the ROC curve. In Section 2 we discuss approaches to deriving the combination score and propose in particular the empirical area under the ROC curve as an objective function of data on which to base the derivation. In Section 3 we show, using data from a protein mass spectrometry experiment, that this approach can yield better classification scores than that based on a likelihood objective function. Simulation studies described in Section 4 indicate that when the logistic regression model holds, the AUC approach is almost fully efficient. We conclude therefore that the AUC is generally not worse and sometimes a lot better than the likelihood for deriving a combination score of multiple predictors. We close in Section 5 with conclusions to date and ideas for further work.

2. Deriving a Combination Score

2.1 *Linear Scores*

We will consider linear scores of the form

$$L_\beta(Y) = Y_1 + \beta_2 Y_2 + \ldots + \beta_P Y_P \quad . \tag{1}$$

Since the component predictors may be functions of the raw predictor data, including transformations and interactions, the linear predictor class is in fact quite large. It includes smoothing and regression splines, kernel methods, generalized additive models, discriminant

scores, support vector machines, and so forth (Hastie, Tibshirani, and Freidman 2001). However, it does exclude one important set of classifiers, namely classic decision trees.

Observe that the linear score does not include an intercept and that the coefficient associated with Y_1 is 1. This is not a restriction since with $\alpha_1 > 0$ (and we can redefine Y_1 as $-Y_1$ to ensure $\alpha_1 > 0$), rules based on the linear predictor $L_\alpha(Y) = \alpha_0 + \alpha_1 L_\beta(Y)$ exceeding a threshold are equivalent to rules based on $L_\beta(Y)$ exceeding a threshold. The ROC curves for $L_\beta(Y)$ and $L_\alpha(Y)$ are the same, so it is enough to consider $L_\beta(Y)$.

Under what circumstances is $L_\beta(Y)$ the "right" combination score for classification to $D = 1$ or 0 based on Y? If the *risk score* is some monotone increasing function of $L_\beta(Y)$,

$$P[D = 1|Y] = g(Y_1 + \beta_2 Y_2 + \ldots \beta_P Y_P) = g(L_\beta(y)) \ , \tag{2}$$

it follows from the Neyman-Pearson lemma (Neyman and Pearson 1933) that rules based on $L_\beta(Y) > c$ are optimal. They are optimal in the sense that no other classification rule based on Y can have even a single accuracy point (FPR, TPR) that lies above the ROC curve for $L_\beta(Y)$. Thus for a fixed FPR, the TPR of the rule $L_\beta(Y) > c$ is higher than the TPR of any other rule with the same FPR. Similarly for fixed TPR, the rule $L_\beta(Y) > c$ has lowest FPR among all rules based on Y with the same TPR. This is an incredibly powerful result that has long been known in the signal detection theory literature (Green and Swets, 1966) but has only recently been highlighted in the statistical literature (McIntosh and Pepe 2002). See also Eguchi and Copas (2002) and Baker (2000) who noted this optimality property for the likelihood ratio function, which is itself a monotone function of the risk score. As a corollary to the optimality of the ROC curve for $L_\beta(Y)$ across its entire domain, it can be shown that rules of form $L_\beta(Y) > c$ minimize the overall misclassification rate and minimize the expected cost of false-negative and false-positive errors combined (Pepe 2003, page 269).

Bayesians have long promoted the risk score function because of these latter two properties. However, optimality of the risk score or of monotone transformations of it is more general and does not require a Bayesian decision theoretic formulation (McIntosh and Pepe, 2002).

In this paper we suppose that the predictors $(Y_1, Y_2, \ldots Y_P)$ are given and the statistical problem is to estimate $\beta = (\beta_2, \ldots, \beta_P)$ from data. We seek estimators that are consistent under the risk score model (2), since under that model $L_\beta(Y)$ is the optimal combination. In addition, we favor procedures that yield linear scores with good classification performance even when the risk score model does not hold. Finally, we seek procedures that allow sampling to depend on the binary outcome D so that case-control studies are accommodated.

2.2 *Objective Functions*

Suppose that we have data for n_D observations truly classified as $D = 1$ and for $n_{\bar{D}}$ with true class $D = 0$. We write the data as $\{Y_{D1}, \ldots Y_{Dn_D}\}$ and $\{Y_{\bar{D}1}, \ldots, Y_{\bar{D}n_{\bar{D}}}\}$ and note that sampling may or may not depend on D. Logistic regression is popular for designs where sampling depends on D because regression parameters other than the intercept can be estimated consistently from the simple prospective log likelihood

$$\log \mathcal{L} = \sum_{i=1}^{n_D} \log P(D_i = 1|Y_{Di}) + \sum_{j=1}^{n_{\bar{D}}} \log P(D_j = 0|Y_{\bar{D}j})$$

even when sampling is retrospective (Prentice and Pyke, 1979). That is, if we assume the risk score model

$$\text{logit}P[D = 1|Y] = \alpha_0 + \alpha_1 Y_1 + \ldots + \alpha_P Y_P \ , \tag{3}$$

the parameters $(\alpha_1, \ldots \alpha_p)$ can be estimated by maximizing

$$\log \mathcal{L}^L(x) = \sum_{i=1}^{n_D} \alpha Y_{Di} - \sum_{k=1}^{n_D+n_{\bar{D}}} \log(1 + e^{\alpha Y_k})$$

where $\alpha Y = \alpha_0 + \alpha_1 Y_1 + \ldots + \alpha_P Y_P$.

The logistic model is a special case of the general linear model (2) with $g(x) = \text{logit}^{-1}(\alpha_0 + \alpha_1 x)$. If we assume that the logistic model holds and calculate the maximum likelihood estimates $(\widehat{\alpha}_1^L, \ldots \widehat{\alpha}_P^L)$, this yields maximum likelihood estimates of $(\beta_2, \ldots \beta_P)$, namely $\widehat{\beta}_p^L = \widehat{\alpha}_p^L / \widehat{\alpha}_1^L$. In summary, the logistic likelihood can be used as an objective function to derive a linear predictor $L_{\widehat{\beta}^L}(Y) = Y_1 + \widehat{\beta}_2^L Y_2 + \ldots + \widehat{\beta}_P^L Y_P$ although consistency is not guaranteed unless the logistic model (3) holds.

Another approach is motivated as follows, assuming only the generalized linear model (2). Optimality of $L_\beta(Y)$ implies that the ROC curve for any other function of Y cannot be higher at any point than the ROC curve for $L_\beta(Y)$. Since $L_\beta(Y)$ has the best ROC curve among all functions of Y, it certainly has the best ROC curve among all linear predictors of the form $L_b(Y) = Y_1 + b_2 Y_2 + \ldots + b_P Y_P$. The idea is to select choices of coefficients $(b_2, \ldots, b_P)$ that yield the best empirical ROC curve for $\{L_b(Y_{Di}), i = 1, \ldots n_D \; ; \; L_b(Y_{\bar{D}j}), j = 1, \ldots n_{\bar{D}}\}$. These are then interpreted as estimates of $(\beta_2, \ldots, \beta_P)$.

The area under the ROC curve (AUC) is the most popular ROC summary index. Interestingly, it can be interpreted as the probability that, for a random case-control pair, the score for the case exceeds that of the control, $P(L_b(Y_{Di}) > L_b(Y_{\bar{D}j}))$.The optimal ROC curve has maximum AUC, so we can use it as the basis for an objective function of the data to estimate β. It is easy to show that the AUC of the empirical ROC curve is the Mann-Whitney U statistic

$$\widehat{\text{AUC}}(b) = \frac{\sum_{i=1}^{n_D} \sum_{j=1}^{n_{\bar{D}}} I\left[L_b(Y_{Di}) > L_b(Y_{\bar{D}j})\right]}{n_D n_{\bar{D}}}.$$

We write the corresponding AUC based estimator of β as

$$\widehat{\beta}^{\text{AUC}} = \text{argmax}(\widehat{\text{AUC}}(b)).$$

Interestingly, this can be recognized as a special case of the maximum rank correlation

estimator of β described by Han (1987). The estimator is known to be consistent and asymptotically normal under the generalized linear model (2). See Sherman (1993) for these results.

2.3 *Relative merits theoretically*

One major attribute of the AUC approach is that it does not require that the link function, g, be specified. It works regardless of the form of the true link function g. On the other hand, the logistic approach depends on the assumption that g is logistic and presumably might fail when g is not logistic.

The logistic model is popular over other forms for g, in part because it accommodates either prospective or retrospective (case-control) designs. Interestingly, we see that the AUC approach shares this property. Because the AUC conditions on the binary response variables $\{D_i = 1, i = 1, \ldots n_D; D_j = 0, j = 1, \ldots n_{\bar{D}}\}$, it allows sampling to depend on D. Thus, it accommodates case-control designs too. Moreover, unlike logistic regression, it does so without restricting the form of the link function to be logistic.

Finally, and most importantly, consider the two approaches when the generalized linear model for the risk score (2) does not hold. The AUC approach still yields a sensible entity, namely the linear combination, $L_b(Y)$, that maximizes the area under the ROC curve (Pepe and Thompson 2000). Even though the resulting linear predictor may not have the optimal ROC curve associated with the risk score, in large samples it optimizes the AUC among all linear combinations of the predictors. In contrast, there are no obvious optimality properties for the linear predictor derived from the logistic likelihood when (2) fails in general.

The one theoretical advantage of the logistic approach over the AUC approach is its asymptotic efficiency when (2) holds and g is logistic. In Section 4 we assess the relative

efficiency of the methods in this setting. First we investigate if the flexibility of the AUC approach offered by the properties of not requiring specification of a link function and having validity when (2) fails, translate into practically meaningful benefits.

3. Applications

3.1 *Protein biomarkers for prostate cancer*

Yasui et al. (2002) describe protein mass spectrometry data generated from the serum of 167 men with prostate cancer and 81 men without cancer. This is one of the cancer biomarker discovery projects being conducted in collaboration with the Early Detection Research Network (Srivastava 1999). After extensive preprocessing of the raw data (see Yasui et al. 2002 for details), the data for analysis comprises of the protein intensity levels at each of 957 mass/charge locations on the protein profile spectrum. Thus there are 957 biomarkers available for predicting prostate cancer status in this case-control study. Most of the biomarkers in this dataset are not predictive of cancer status. Only 144/957(15%) empirical AUCs exceeded 75%. Yasui et al. used stepwise logistic regression to derive a linear combination score that could be used to classify subjects as having prostate cancer or not based on serum protein mass spectrometry.

The issues involved in predictor selection, particularly from high-dimensional data, are beyond the scope of the current paper. We used the data to simply illustrate that different choices of objective function, logistic likelihood versus AUC, can lead to linear predictors with substantially different classification performance. For simplicity and ease of illustration, combinations of only 2 biomarkers were considered.

3.2 *AUC versus MLL Combinations*

For the most part linear combinations that maximized the logistic likelihood had similar performance in terms of their ROCs to those that maximized the empirical AUC. However there were a variety of combinations where the AUC-based linear combination performed substantially better. Figure 1 shows some examples. The lines on the scatterplots show the directions of the linear scores, $Y_1 + \beta_2^L Y_2 =$ constant and $Y_1 + \beta_2^{\text{AUC}} Y_2 =$ constant. Recall that decision rules based on the linear scores classify a subject as a case if their linear score exceeds a constant, i.e., lies above the line $Y_1 + \beta_2 Y_2 =$ constant. Different choices of constant yield different lines parallel to those shown. Each line has an associated (FPR, TPR) point on the ROC curve shown in the right hand panels of Figure 1. Note that these scatterplots and ROC curves do not incorporate sampling variability. The joint distribution displayed in the left panel essentially provides a statistical simulation model and the slopes $(\beta_2^L, \beta_2^{\text{AUC}})$ can be regarded as large sample values of the estimates $\widehat{\beta}_2^L$ and $\widehat{\beta}_2^{\text{AUC}}$. Similarly, the ROC curves are calculated using the joint marker distribution displayed, which for now is assumed to be the true distribution, not a sample from some underlying truth. Sampling variability is addressed later in Section 4.

In each of the four examples shown, the ROC curve for $Y_1 + \beta_2^L Y_2$ is only slightly better than that for the better of the two component markers while the ROC curve for $Y_1 + \beta_2^{\text{AUC}} Y_2$ is clearly superior. Table 1 shows the areas under these ROC curves.

By definition the ROC curve for $Y_1 + \beta_2^{\text{AUC}} Y_2$ is superior to that of each of the component markers. However, there is no such guarantee for the logistic likelihood based combination. Figure 2 and Table 1 lower panel display some marker combinations where $Y_1 + \beta_2^L Y_2$ has *poorer* classification performance than that of a single component marker. This raises seri-

ous concerns about the use of the logistic likelihood in general to derive a linear score for classification in settings where the logistic model fails. Although the likelihood is increased by using it to combine markers (Table 1), the operating characteristics of the combination for classification may deteriorate substantially relative to using a single marker.

4. Finite Sample Simulations

4.1 *Logistic Model*

Under the logistic model, $\text{logit}P[D = 1|Y] = \alpha_0 + \alpha_1(Y_1 + \beta_2 Y_2), L_\beta(Y) = Y_1 + \beta_2 Y_2$ is the optimal combination in the sense that any other combination of Y_1 and Y_2 cannot have an (FPR, TPR) point that lies above the ROC curve for $Y_1 + \beta_2 Y_2$. Since $\widehat{\beta}_2^L$ and $\widehat{\beta}_2^{\text{AUC}}$ are both consistent estimates of β_2, in large samples the logistic likelihood and AUC approaches both yield the optimal linear combination (clearly the logistic model does not hold in any of the scenarios shown in Figures 1 and 2) since $\widehat{\beta}_2^L$ is not optimal. In small samples, however, the estimates will differ (Figure 3). Given that $\widehat{\beta}_2^L$ is the statistically efficient estimator of β_2 under the logistic model, one would expect that it would on average yield the better linear combination function based on a finite sample of data from the logistic model. To investigate this we simulated samples of $n_D = 50$ cases and $n_{\bar{D}} = 50$ controls with bivariate normal marker distributions (Y_1, Y_2), having covariance matrix the identity and mean vectors $(\mu_{D1} = 1, \mu_{D2} = 1)$ in cases and $(\mu_{\bar{D}1} = 0, \mu_{\bar{D}2} = 0)$ in controls. This configuration induces the logistic model: $\text{logit}\{P(D = 1|Y_1, Y_2)\} = -1 + Y_1 + Y_2$. Thus $\beta_2 = 1.0$ in this setting.

The top row of Table 2 and Figure 3 show the results of 500 simulations. As expected both $\widehat{\beta}_2^L$ and $\widehat{\beta}_2^{\text{AUC}}$ have little bias and $\widehat{\beta}_2^L$ is more efficient than $\widehat{\beta}_2^{\text{AUC}}$ $(\text{var}(\widehat{\beta}^L < \text{var}(\widehat{\beta}^{\text{AUC}}))$. To gauge the performances of the estimated linear combination functions, $Y_1 + \widehat{\beta}_2^L Y_2$ and $Y_1 + \widehat{\beta}_2^{\text{AUC}} Y_2$, we calculated the corresponding AUCs using the true underlying logistic model.

Results displayed in Table 2 and Figure 3 indicate that the classification accuracies of the linear scores are very similar, the average AUC being 0.838 for both the logistic approach and the AUC-based method. Thus, despite the fact that $\widehat{\beta}_2^L$ is a more efficient estimator of β_2, it does not appear to yield substantially better classification performance than $\widehat{\beta}_2^{\text{AUC}}$ under the logistic model. Qualitatively similar conclusions are found for alternative choices of means in the bivariate normal marker model (Table 2).

4.2 *Protein Marker Models*

We simulated data from the configurations depicted in the scatterplots of Figure 1 using the sample sizes enrolled in the protein mass spectrometry study. In essence, this is bootstrap resampling. Table 3 summarizes distributions of $\widehat{\beta}_2^L$ and $\widehat{\beta}_2^{\text{AUC}}$ along with the AUCs of the associated linear combinations of markers. Note that the underlying true marker distribution shown in Figure 1 is used to calculate the AUC, not the bootstrapped data, which would only estimate the AUCs.

The sampling distributions of the estimates around their asymptotic values are shown in Figure 4 for scenario 1. Clearly in small samples the AUC-based approach yields linear combinations with superior classification performance. Moreover, the AUC of $Y_1 + \widehat{\beta}_2^{\text{AUC}} Y_2$ is very close to the best possible value, $\text{AUC}(\beta_2^{\text{AUC}})$ even with sample sizes of $n_D = 167$ and $n_{\bar{D}} = 81$. Ninety percent of the $\text{AUC}(\widehat{\beta}_2^{\text{AUC}})$ values are above 75%. This suggests that the sample sizes are adequate to develop a classifier that combines v426 and v427 linearly. Such considerations could be used as the basis for sample size calculations in the design of studies to combine markers.

5. Discussion

The main contribution of the current paper is to demonstrate that the choice of objective function to be optimized is crucial to deriving a linear combination of markers for classification. If classification performance is measured with the area under the ROC curve then one should use it to generate the linear function. We showed with real data that maximizing the logistic likelihood (also called the entropy (Hastie, Tibshirani and Friedman 2001)) can yield unacceptably poor classification performance. We note however that likelihood-based regression methods are frequently employed in practice to combine markers for classification (Hastie, Tibshirani and Freidman 2001). These are appropriate and statistically efficient if the regression model is correct, but our results indicate that they can behave dismally otherwise. Hastie, Tibshirani and Friedman (2001) note that regression models are difficult to verify in higher dimensions. Therefore the use of likelihood-based methods in higher dimensions may be particularly problematic.

We have focused on the AUC objective function here which is estimated non-parametrically with the Mann-Whitney U-statistic. It accommodates case-control data, yields the optimal linear combination asymptotically under the generalized linear model (2) and most importantly is well motivated by considerations of classification accuracy even when the generalized linear model does not hold. We have previously proposed use of the AUC to combine markers (Pepe and Thompson 2000) but did not note its optimality properties under the generalized linear model or demonstrate its superiority to likelihood methods when the model fails. In addition, we also note here that the estimate obtained by maximizing the empirical AUC can be viewed as a special case of the maximum rank correlation estimator for which general asymptotic distribution theory has been developed. Eguchi and Copas (2002) also discuss

the use of the AUC for deriving linear scores, although their approach to calculating the AUC is complex and they only demonstrate its superiority to logistic regression with one hypothetical pathologic example.

Objective functions other than the AUC might also be considered for developing a linear classification score. We have suggested the *partial* AUC (Pepe and Thompson 2000) that restricts attention to a region of the marker space associated with practically relevant (FPR, TPR) points (see also McIntosh and Pepe 2002). The misclassification rate (MCR) associated with Bayes' rule is another natural objective function to consider (Hastie, Tibshirani and Friedman 2001, p27). However, it depends on the ratio of cases to controls in the sample, which in a case-control study will differ from the ratio in the populations. Minimizing the study MCR may produce a linear score that does not minimize the population MCR. Weighting the misclassification probabilities according to the population prevalence of cases should perhaps be considered.

Although we have made a case for optimizing the AUC to combine markers, much work remains to be done before the approach can be routinely applied in practice. Computational algorithms need to accommodate the fact that the empirical AUC is not a continuous function. We dealt with only two markers in our applications and used a simple search routine for optimization. More sophisticated approaches would be required when the number of markers exceeds 2. Consideration of multiple markers also highlights the need for marker selection algorithms. We have suggested a simple stepwise algorithm (Pepe and Thompson 2000), but a more rigorous development is warranted.

We have discussed maximizing the AUC to derive a linear score for the purposes of classification. The AUC can also be motivated simply as a technique for fitting a regression

model. It is a technique that is robust to the choice of link function g because it does not require that the link even be specified. Although this would appear to be an advantage over logistic regression, Li and Duan (1989) and our own simulation studies (not shown) indicate that logistic regression is itself quite robust. That is, under certain conditions stated in Li and Duan (1989), logistic regression performs well even when the link function is not logistic. However, in practice there is no guarantee that the conditions will be met, so maximizing the AUC may still be preferred for fitting the generalized linear model $P(D = 1|Y) = g(L_\beta(Y))$.

Acknowledgements

We would like to thank Noelle Noble for assistance with preparing the manuscript and the referees and Holly Janes for helpful comments on an earlier version of the paper.

Résumé

References

Baker, S.G. (2000). Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics* **56**, 1082–1087.

Baker, S. G. (2003). 'The Central Role of Receiver Operating Characteristic (ROC) Curves in Evaluating Tests for the Early Detection of Cancer. *Journal of the National Cancer Institute*, **95**, 511–515.

Copas, J.B. and Corbett, P (2002). Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika* **89**(2), 315–331.

Eguchi, S. and Copas, J.B. (2002). A class of logistic-type discriminant functions.

Biometrika **89**(1), 1–22.

Green, D.M. and Swets, J.A. (1966). *Signal detection theory and psychophysics.* Wiley, New York.

Han, A.K. (1987). Non-parametric analysis of a generalized regression model. The maximum rank correlation estimator. *Journal of Economics* **35** 303–316.

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

Li, K-C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics* **17** 1009–1052.

McIntosh, M.S., Pepe, M.S. (2002). Combining several screening tests: optimality of the risk score. *Biometrics* **58** 657–664.

Neyman, J. and Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypothesis. *Philosophical Transactions of the Royal Society of London, Series A* **231** 289–337.

Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction* Oxford University Press, United Kingdom.

Pepe, M.S. and Thompson M.L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1**(2) 123–140.

Pepe, M.S., Etzioni R., Feng Z., Potter J.D., et al. (2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*

93(14), 1054–1061.

Pepe, M.S., Janes, H., Longton, G., Leisenring, W., et al. (2004). Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker. *American Journal of Epidemiology* **159**(9), 882–890.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–412.

Sherman R.P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrics* **61** 123–137.

Srivastava, S. (1999). Early Detection Research Network. *Disease Markers* **15** 213–219.

Yasui, Y., McLerran, D., Adam, B-L., Winget, et al. (2003). An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *Journal of Biomedicine and Biotechnology* **4** 242–248.

Zhou, X-H., Obuchowski, N.A., McClish, D.K. (2002) *Statistical Methods in Diagnostic Medicine* Wiley, New York.

Table 1

Results for selected marker pairs. See Figures 1 and 2 for related results.

		AUC				$\log\mathcal{L}$		
Marker 1	Marker 2	Marker 1	Marker 2	$L_{\widehat{\beta}_2^L}$	$L_{\widehat{\beta}_2^{\mathrm{AUC}}}$	Marker 1	Marker 2	$\widehat{\beta}_2^L$
v426	v427	.54	.58	.62	.76	−156.6	−156.0	−155.9
v652	v877	.71	.72	.72	.84	−141.0	−139.3	−139.3
v653	v831	.70	.73	.73	.88	−141.7	−139.4	−139.4
v653	v741	.70	.73	.73	.86	−141.7	−139.5	−139.5
v354	v365	.68	.57	.51	.70	−156.6	−156.3	−155.6
v182	v530	.51	.69	.51	.70	−156.5	−156.3	−155.9
v30	v93	.54	.64	.50	.68	−156.6	−156.4	−156.3
v509	v637	.63	.71	.63	.81	−150.3	−153.9	−150.3

Table 2

Results of 500 simulations from a logistic regression model with 50 cases and 50 controls.

				$\widehat{\beta}_2^L$		$\widehat{\beta}_2^{\text{AUC}}$		$\text{AUC}(\widehat{\beta}_2^L)$		$\text{AUC}(\widehat{\beta}_2^{\text{AUC}})$	
(μ_{D1}, μ_{D2})		β_2	$\text{AUC}(\beta_2)$	(mean, sd)		(mean, sd)		(mean, sd)		(mean, sd)	
1.0	1.0	1.0	0.842	1.033	0.388	1.048	0.443	0.838	0.005	0.838	0.006
1.0	2.0	2.0	0.943	2.311	1.161	2.805	4.839	0.941	0.002	0.941	0.004
2.0	2.0	1.0	0.977	1.028	0.315	1.087	0.400	0.975	0.002	0.975	0.002
0.5	0.5	1.0	0.693	1.308	1.270	1.668	2.752	0.684	0.012	0.682	0.014

Table 3

Results of 1000 bootstrap samples from four marker combinations where maximizing AUC outperforms maximizing the logistic likelihood.

scenario	marker		β_2^L	$\widehat{\beta}_2^L$ (mean, sd)		AUC(β_2^L)	AUC($\widehat{\beta}_2^L$) (mean, sd)	
	1	2						
1	426	427	2.487	6.050	91.8556	0.619	0.636	0.0853
2	652	877	40.330	8.612	39.9891	0.723	0.737	0.0307
3	653	831	152.108	9.196	63.5295	0.728	0.744	0.0379
4	653	741	33.294	9.987	82.0671	0.726	0.740	0.0365

scenario	marker		β_2^{AUC}	$\widehat{\beta}_2^{\mathrm{AUC}}$ (mean, sd)		AUC(β_2^{AUC})	AUC($\widehat{\beta}_2^{\mathrm{AUC}}$) (mean, sd)	
	1	2						
1	426	427	1.088	1.092	0.0375	0.756	0.754	0.0038
2	652	877	1.024	1.029	0.0152	0.844	0.843	0.0027
3	653	831	1.028	1.026	0.0114	0.878	0.876	0.0021
4	653	741	1.012	1.023	0.0185	0.864	0.862	0.0014

FIGURE LEGENDS

Figure 1. Scenarios where maximizing the AUC yields a substantially better linear classification score than maximizing the logistic likelihood. Scatterplots show data for cases (•) and controls (∘). Lines and ROC curves for the scores derived from the likelihood (dashed) and AUC (solid) objective functions are shown. See Table 1 for related results.

Figure 2. Scenarios where maximizing the logistic likelihood yields a linear combination with AUC worse than that of a single marker. See Table 1 for related results. The direction of the logistic likelihood fitted linear combination is shown (dashed) along with data points for cases (•) and controls (∘). ROC curves are also displayed.

Figure 3. Scatterplots of $\widehat{\beta}_2^L$ versus $\widehat{\beta}_2^{\mathrm{AUC}}$ and $\mathrm{AUC}(\widehat{\beta}_2^L)$ versus $\mathrm{AUC}(\widehat{\beta}_2^{\mathrm{AUC}})$ for simulated data from the logistic model with independent normally distributed predictors (top row of Table 2).

Figure 4. Distributions of $\widehat{\beta}_2^L$ and $\widehat{\beta}_2^{\mathrm{AUC}}$ and of $\mathrm{AUC}(\widehat{\beta}_2^L)$ and $\mathrm{AUC}(\widehat{\beta}_2^{\mathrm{AUC}})$ for data generated by bootstrap resampling of markers v426 and v427 in the protein mass spectrometry dataset. Kernal density estimates of the distributions are truncated at the minimum and maximum values.

Fig. 1

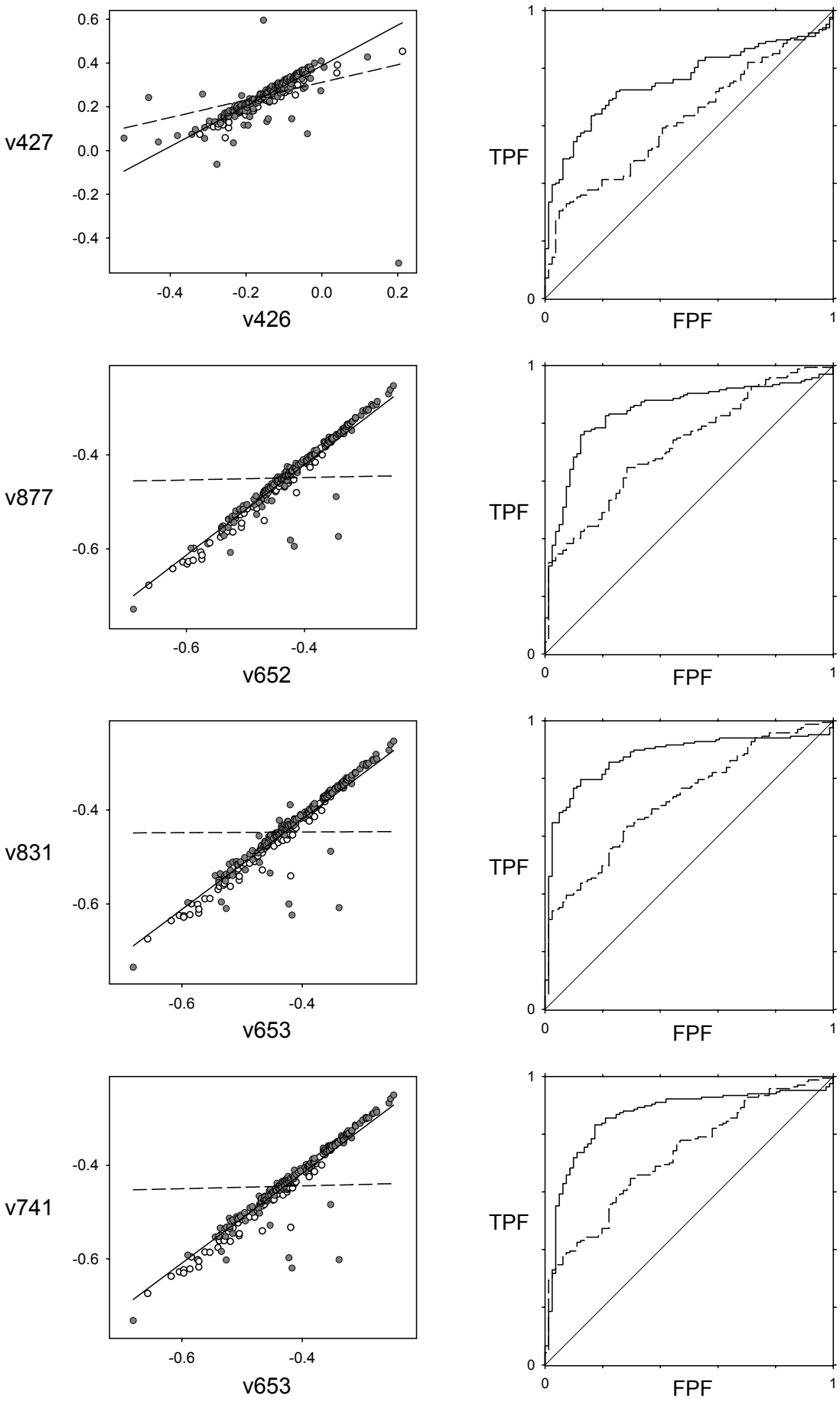


Fig. 2

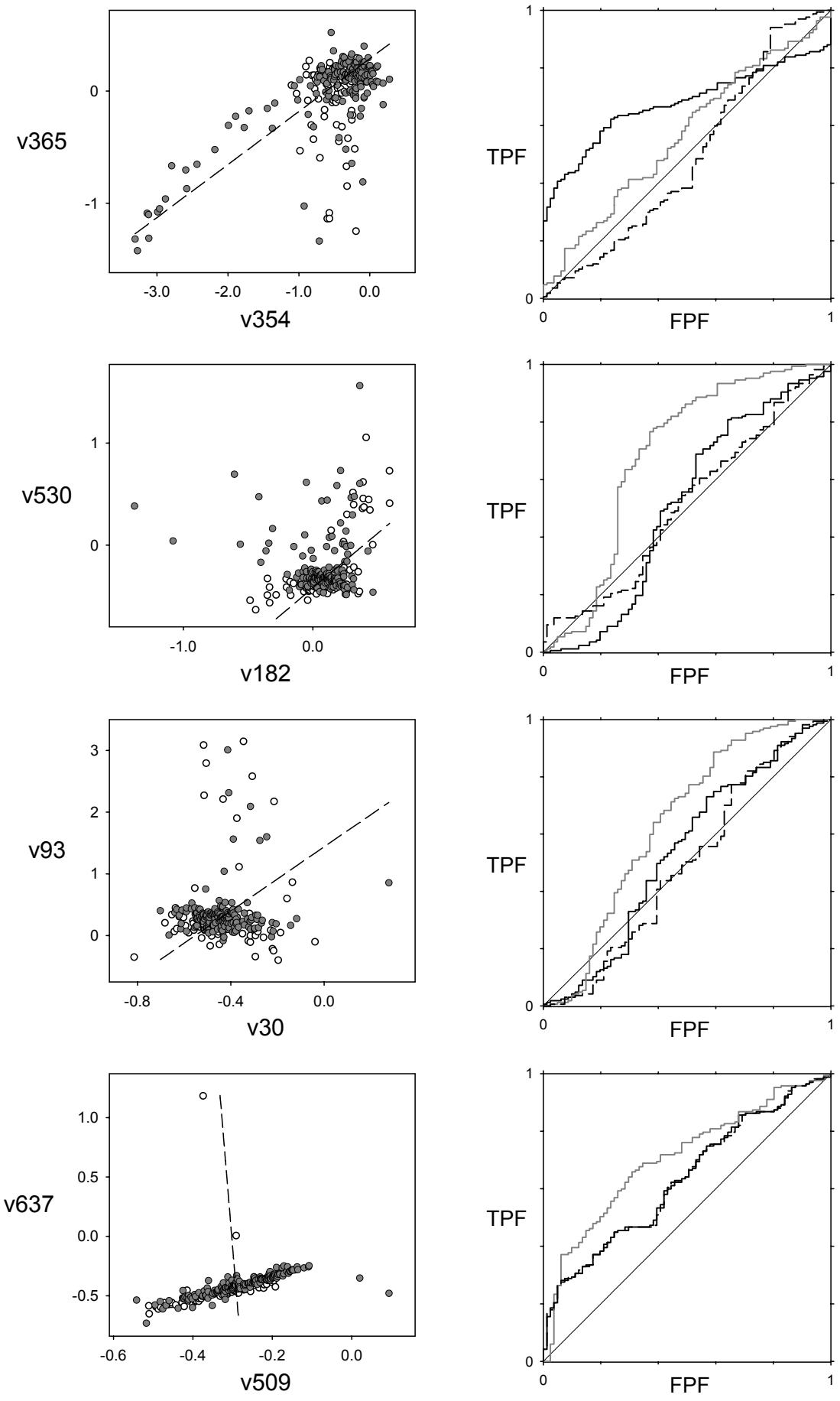
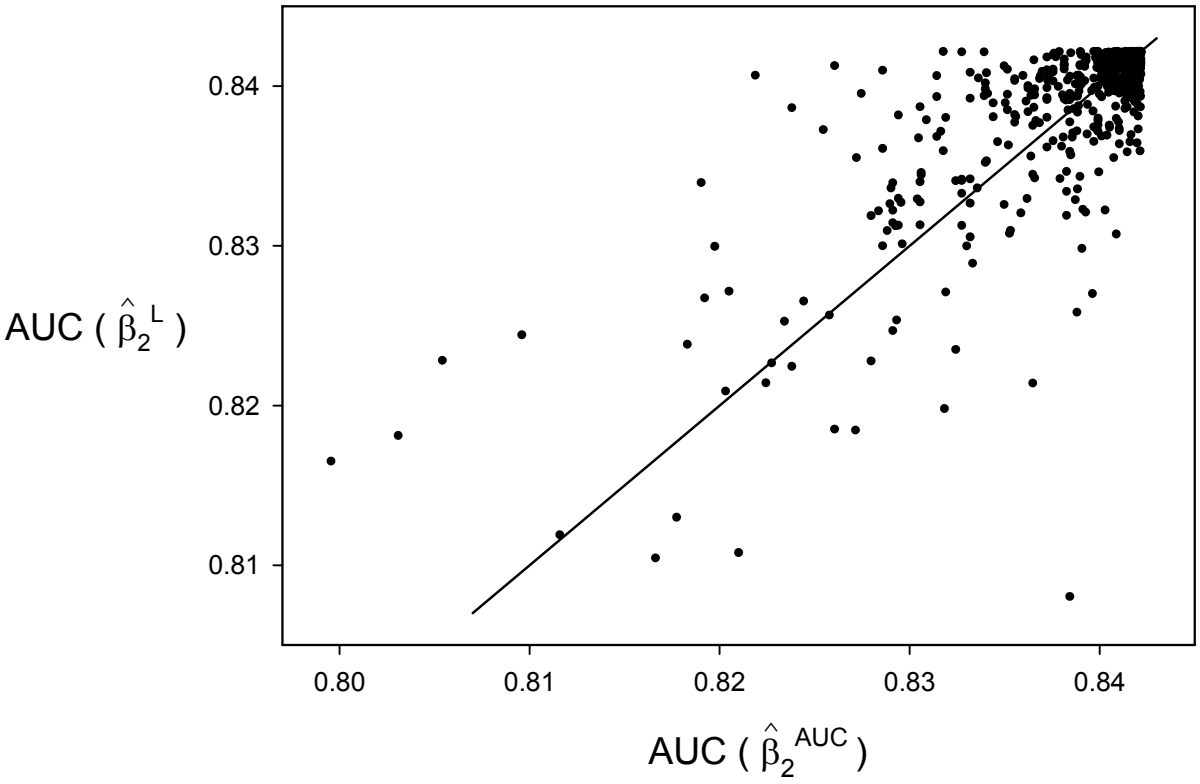


Fig. 3



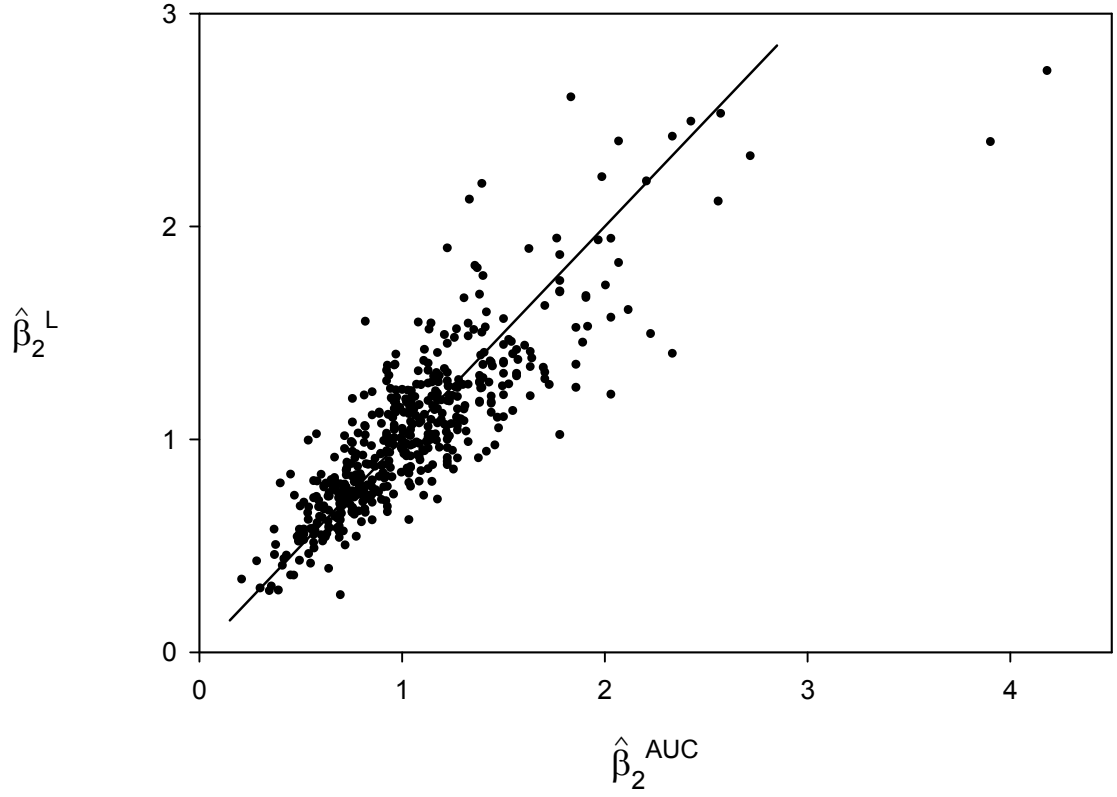
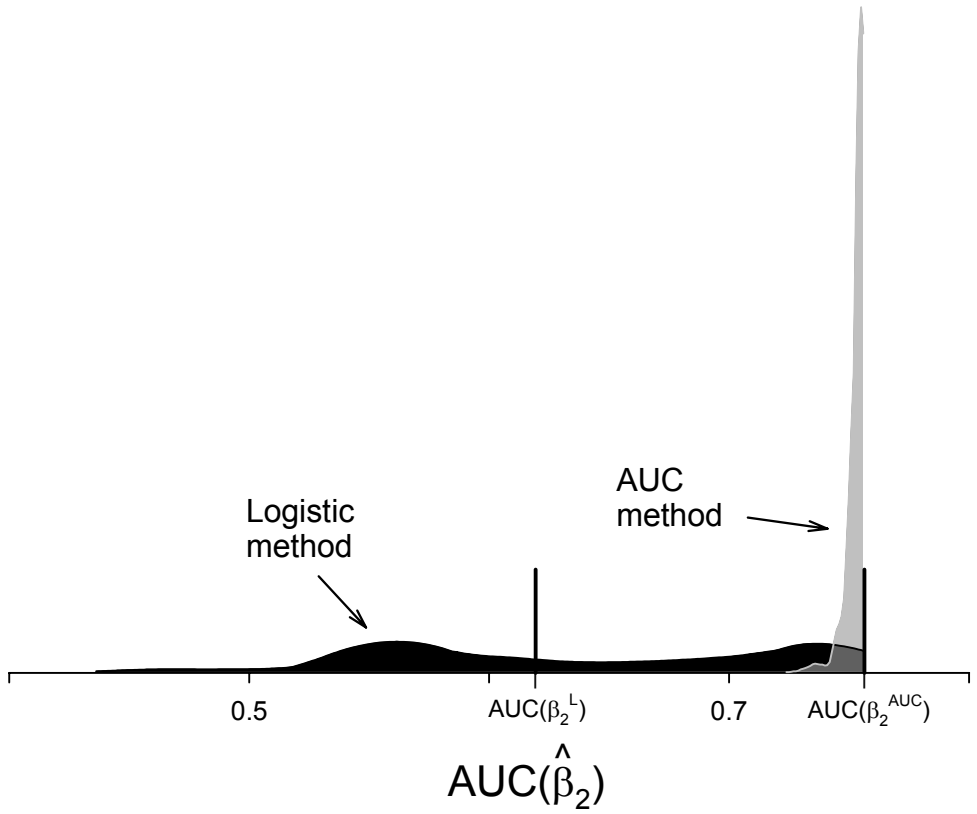
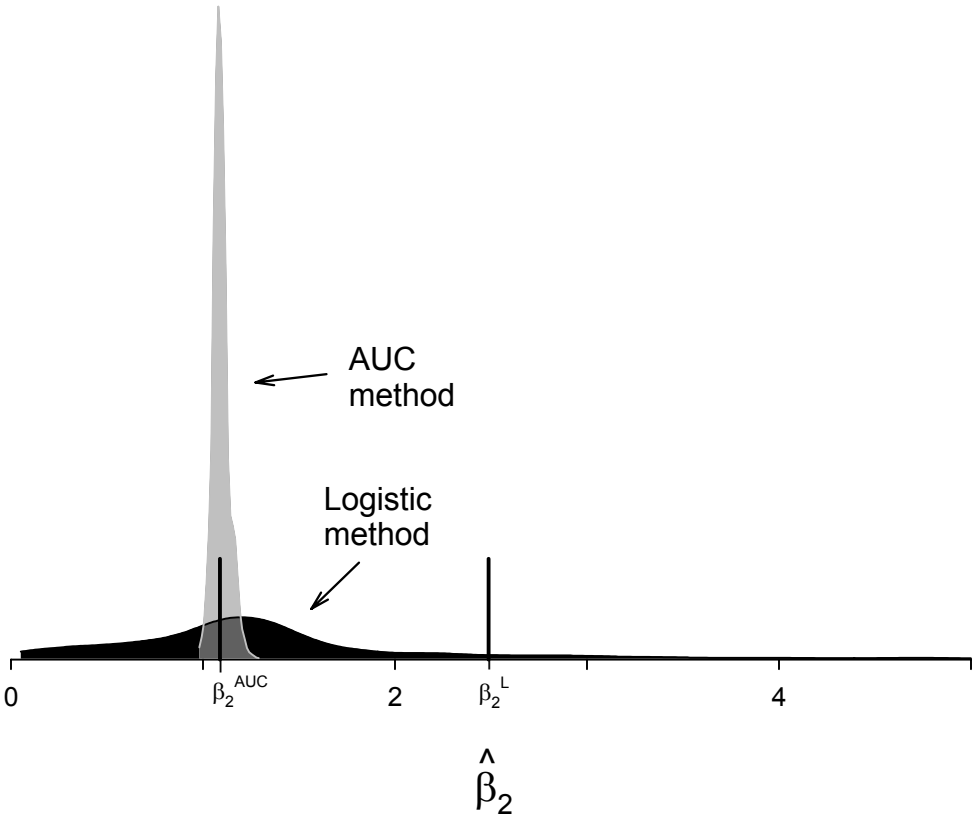


Fig. 4


Logistic
method
AUC
method
0.5
AUC(β2 L)
0.7
AUC(β2 AUC)
AUC(β̂2)


AUC
method
Logistic
method
0
β2 AUC
2
β2 L
4
β̂2