Statistics and Its Interface Volume 7 (2014) 69–73

Inherent difficulties in nonparametric estimation of the cumulative distribution function using observations measured with error: Application to high-dimensional microarray data

George W. Wright, Lori E. Dodd, and Edward L. Korn*

Distribution function estimation is important in many biological applications. A very simple example is given to show that with the addition of normal errors, data from very different underlying distributions can generate nearly identical distributions of observations. Therefore, in some situations it can be essentially impossible to accurately estimate an underlying cumulative distribution function from a reasonable number of observations measured with error. An application is given involving estimating the distribution function of differential gene expression based on more than fifty thousand genes.

AMS 2010 subject classifications: Primary 62C12; secondary 62G07; 62P10.
Keywords and phrases: Empirical Bayes, Stability, Microarray Data, Mixture Models, Measurement Error, Shrinkage.

1. INTRODUCTION

A common biometrical problem is to estimate the distribution of a random quantity based on an independent and identically distributed random sample $X_1, X_2, \ldots, X_n$. The empirical distribution function $\hat{F}_X(x) = \sum I(X_i \leq x)/n$, where $I()$ is the 0-1 indicator function, is an excellent estimator in that it unbiasedly estimates the underlying cumulative distribution function $F_X(x)$ and, by the Glivenko-Cantelli theorem, is consistent for it in the sense that

$$P\left(\lim_{n\to\infty} \sup_x |\hat{F}_X(x) - F_X(x)| = 0\right) = 1$$

[1]. The empirical distribution function can also be thought of as the nonparametric maximum likelihood estimator of F_X [19]. Suppose instead of observing $X_1, X_2, \ldots, X_n$, we observe $Y_i = X_i + Z_i$, $i = 1, \ldots, n$, where the Z_i are independent measurement error terms assumed to have known distributions. We may still be interested in estimating F_X in this situation.

One potential application is estimation of the true distribution of a dietary component based on a large sample of individuals with self-reported dietary data [9]. A second is in estimation of the underlying distribution of gene expression difference, which we will focus on here. For example, the solid gray line in Fig. 1 displays the empirical distribution function of 54,675 two-sample t-statistics, representing tests of differential gene expression between two classes of lymphoma ($n_1 = 74$, $n_2 = 77$). This application will be discussed in detail in Section 3, but here we note that the observed t-statistic for gene $i(Y_i)$ can be thought of as measurement error Z_i plus a true standardized differential gene expression $X_i = (\mu_{i1} - \mu_{i2})/\sqrt{\sigma^2(n_1^{-1} + n_2^{-1})}$. Knowing the distribution of true standardized differential gene expressions across the genes would be useful in many applications, such as estimating sample size requirements for developing classifiers based on thousands of genes [5].

If one assumes the distribution of the X's belongs to some parametric family of distributions, then the estimation of F_X from $Y_1, Y_2, \ldots, Y_n$ is straightforward, e.g., using maximum likelihood estimation. A nonparametric approach is more challenging. Assume that Z_i has a normal distribu-

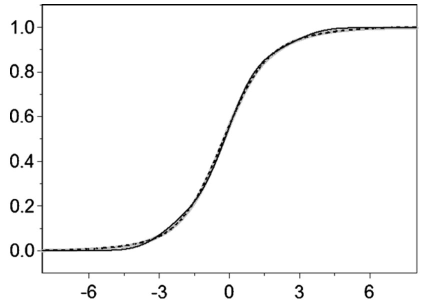

Figure 1. Cumulative distribution graphs of 54,675 observed t-statistics (solid gray), and estimated distribution functions of Y from a two-component normal mixture model (dashed black) and a three-point distribution (solid black). The graphs have been truncated at ±8, omitting 0.33% of the observed t-statistics. The three curves overlap and are virtually indistinguishable.

*Corresponding author.

tion with mean zero and known variance σ_i^2, $i = 1, \ldots, n$. One can estimate the non-parametric maximum likelihood estimator of F_X by maximizing

$$L(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} \left(\int \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2\sigma_i^2}(y_i - \alpha)^2 \right) dF(\alpha) \right)$$

over all possible distribution functions F. This estimator is discrete with at most d mass points, $\tilde{F}_X(x) = \sum_{j=1}^{d} I(x \leq \hat{\alpha}_j)/d$, where d is the number of distinct $Y_1, Y_2, \ldots, Y_n$ [11, 14]. Smoothed nonparametric estimators similar to $\tilde{F}_X$ are also possible, e.g. [12], as well as other nonparametric estimators of F_X, e.g. [7, 16, 20].

The type of deconvolution problem being considered here is known to be hard, in the sense of getting accurate estimates, unless the sample sizes are large [3]. Fortunately, in our application, the sample size is over fifty thousand. We give a simple hypothetical example in the next section to show that unfortunately even in these settings none of these nonparametric approaches is likely to work for estimating F_X. We return to our application in Section 3 to demonstrate that the problems identified in Section 2 occur in a real application. We end with a brief discussion.

2. A SOBERING EXAMPLE

Assume that the sample size was so large that we essentially know the distribution function $F_Y$of Y exactly. We first can ask in this situation if it is possible to calculate F_X; this is an identifiability question. To be specific, suppose the distribution of $X+Z$ is the same as the distribution of $X' + Z$. Does it follow that the distribution of X and X' must be the same? In general, the answer is no. But if Z has a normal distribution, the answer is yes (due to the fact that the characteristic function for the normal distribution is never zero).

More to the point here, suppose Z has a normal distribution and the distribution function of $X + Z$, F_{X+Z}, is not equal to but very close to the distribution function of $X'+Z$, $F_{X'+Z}$; that is, $F_{X'+Z}(t) \cong F_{X+Z}(t)$, for all t. Then does it follow that $F_{X'}(t) \cong F_X(t)$? Unfortunately not, as the following example demonstrates. Let Z and X have standard normal distributions, so that $F_{X+Z}(y) = \Phi(y/\sqrt{2})$, where Φ is the cumulative distribution function of a standard normal distribution. Let X' have a three-point distribution: $\pm\sqrt{3}$ each with probability $1/6$, and 0 with probability $2/3$. Note that the variance and kurtosis of X' are 1 and 0, respectively, so that first four moments of $X + Z$ and $X' + Z$ match. The distribution function of $X' + Z$ is

$$F_{X'+Z}(y) = \frac{1}{6}\Phi(y - \sqrt{3}) + \frac{2}{3}\Phi(y) + \frac{1}{6}\Phi(y + \sqrt{3}),$$

and is plotted in Fig. 2 with the distribution function of $X + Z$. The distribution functions are indistinguishable on the plot; they differ by at most 0.0063 (Fig. 3). On the other hand, the distribution functions of F_X and $F_{X'}$ differ substantially; by as much as 1/3 at $x = 0$ (Fig. 4).

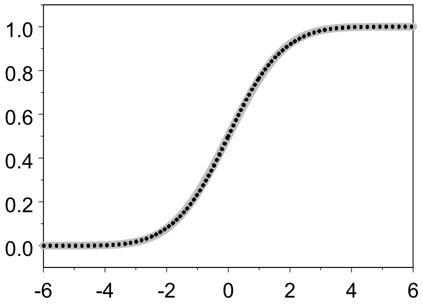

Figure 2. Cumulative distribution graphs of $X + Z$ (solid gray) and $X' + Z$ (dashed black). The two curves overlap and are virtually indistinguishable.

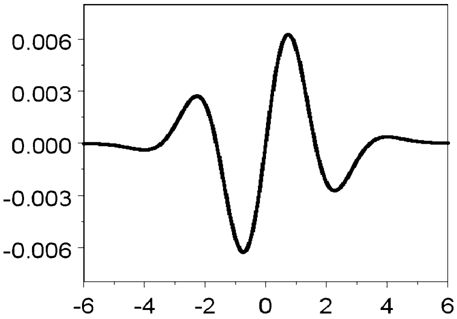

Figure 3. Difference between the cumulative distribution of $X + Z$ and $X + Z'$.

The implications of this example are that, since even with extremely large sample sizes we are unlikely to be able to distinguish distribution functions that look like those in Fig. 1, nonparametric estimation of a distribution function in the presence of known normally distributed measurement error may be infeasible. This is true even though the nonparametric maximum likelihood estimator $\tilde{F}_X$ is consistent for F_X in this situation [8].

As an aside, we note that this example also has implications for the stability of Cramer's theorem, which states that if $Y = V + W$ has a normal distribution, and V and W are independent, then V and W must also be normally distributed [4]. Sapogov [18] showed that if Y has approximately a normal distribution in the sense that there exists aμ and σ^2 such that

$$\sup_y \left| F_Y(y) - \Phi\left(\frac{y-\mu}{\sigma}\right) \right| < \varepsilon,$$

then the difference between the distribution function for V and a normal distribution function could be bounded by $C\sigma_V^{-3}(\log(1/\varepsilon))^{-1/2}$ where $\sigma_V^2$is a truncated variance of V, and C is a universal constant. (See Kagan et al. [10] for

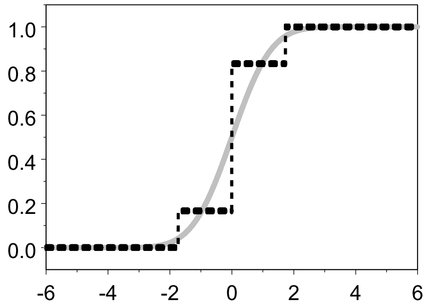

Figure 4. Cumulative distribution of X (solid gray), and X' (dashed black).

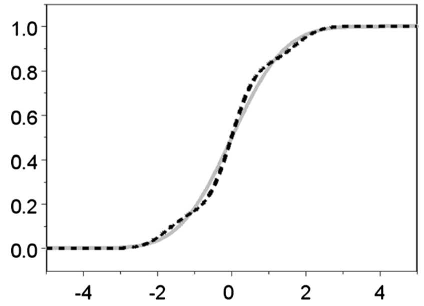

Figure 5. Cumulative distribution graphs of $X + Z/2$(solid gray) and $X' + Z/2$ (dashed black).

details). Maloshevskii [15] showed that asymptotically the Sapogov bound could not be improved. The example in Fig. 2, for which $\varepsilon = 0.0063$, suggests that for small but non-infinitesimal ε the Sapogov bound is unlikely to be useful.

In the example given the measurement error Z was of similar magnitude to the variability of X. If the measurement error is significantly smaller, then one may be able to distinguish different X's from observed $X + Z$. For example, Fig. 5 displays the cumulative distribution functions of $X + Z/2$ and $X' + Z/2$ where Z is a standard normal distribution. Unlike Fig. 2, one can now distinguish the curves. However, note (1) that if the X and X'were not as different as in Fig. 4, or (2) the sample size was not essentially infinite (as implied here by observing the distribution functions of $X + Z/2$ and $X' + Z/2$ exactly), then there may be limited ability to distinguish X and X' based on the empirical distribution function the observations measured with error.

Although a nonparametric estimator of $F_X(x)$ may be substantially off at specific x's (as in Fig. 4), it may still be useful as an input to other procedures. However, these applications will have to be considered on a case by case basis. For example, consider estimation of X_i using nonparametric empirical Bayes estimation via the posterior mean, $\hat{X}_i = E(X_i | X_i + Z_i = y_i; \tilde{F}_X)$. The simple three-point distribution (X') versus normal example (X) shows that this will unlikely be successful for extreme values of y_i, which will typically be the ones of greatest biological interest. In particular, $E(X_i | X_i + Z_i = y_i) = y_i/(1 + \sigma_i^2)$ but, for large y_i, $E(X_i' | X_i' + Z_i = y_i) = \sqrt{3}$. Note that this example suggests possible difficulties for nonparametric empirical Bayes estimation whether F_X is first estimated and then used to estimate the posterior mean, or the posterior mean is estimated directly without first completely estimating F_X as was done originally by Robbins [17]; see Brown [2] and Efron [6] for recent examples of the latter approach.

The example presented also indicates a potential danger even in parametric empirical Bayes settings. If an investigator observes that the data distribution matches well to a given parametric mixture model, he may naively assume that the parametric assumptions used to generate the model were valid. But as we demonstrated, it is possible that even if the distribution data appears to match well to his model, it is possible that the underlying distribution is very different from the modelled one.

3. APPLICATION

Lenz et al. [13] provided a data set of 181 diffuse large B-cell lymphoma samples whose gene expression had been analyzed with a Affymetrix U133 plus 2.0 array containing 54,675 probesets. Of these, 74 were previously identified as being of the Activated B-cell (ABC) subtype, 77 had been classified as of the Germinal center B-cell subtype while 31 were unclassifiable [21]. Although one might be interested in the estimate of the distribution across the 54,675 probesets of the fold change for differential expression between the ABC and GCB lymphoma subtypes, we focus here on the distribution of the t-statistics (and standardized differential gene expression), which might be more relevant for some types of inferences and provides a direct analogy with the hypothetical example of Section 2: With the moderately large (for this type of experiment) number of sample specimens, each t-statistic can be considered as $Y_i = X_i + Z_i$ where $X_i = (\mu_{i1} - \mu_{i2})/\sqrt{\sigma^2(n_1^{-1} + n_2^{-1})}$ and Z_i has an approximate standard normal distribution. The solid gray line in Fig. 1 is the empirical cumulative distribution function of the 54,675 observed t-statistics.

We will now construct two very different distributions (X and X') such that both $X + Z$ and $X' + Z$ have distributions that look very similar to the distribution of the observed t-statistics, where Z has a standard normal distribution. One distribution (X) is a mixture of two normal distribution (a normal with mean -0.02 and variance 7.89 with probability 0.32, and a normal with mean $= -0.28$ and variance $= 0.71$ with probability 0.68) and the other (X') is a three-point distribution (-2.64 with probability 0.19, -0.06 with probability 0.69, and 2.82 with probability 0.12). The cumulative distribution functions for these two distributions are displayed in Fig. 6. The distributions of $X + Z$ and $X' + Z$ are displayed in Fig. 1, and are very close to the empirical distribution function of the observed

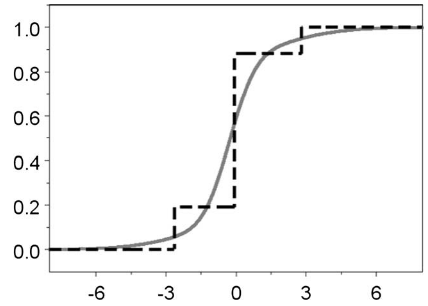

Figure 6. Cumulative distribution of two-component normal mixture X (solid gray), and three-point distribution X' (dashed black) for standardized differential gene expression for 54,675 genes.

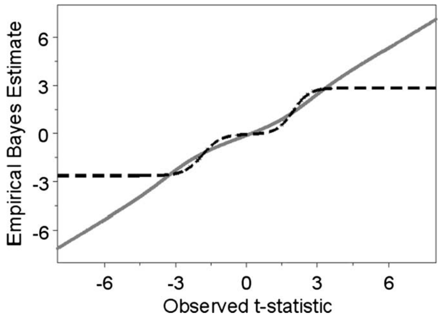

Figure 7. Empirical Bayes shrinkage estimator as a function of the observed t-statistic using a two-component normal mixture for X (solid gray), and three-point distribution for X (dashed black) for standardized differential gene expression for 54,675 genes.

t-statistics, the largest difference between the modelled and observed empirical distribution functions being 0.006 and 0.014 respectively. So either X or X' would be a reasonable estimate of the true underlying distribution. However in terms of the Bayes shrinkage we discussed in Section 2, these two models show markedly different results (Fig. 7). Thus it is difficult to say what the correct empirical Bayes shrinkage estimate should be in this case. Note that we are not claiming that the models X and X' reflect the true underlying biology but rather we present them as extreme examples to highlight the potential variability of empirical Bayes estimation.

4. DISCUSSION

The example and application presented in this paper are meant to provide a caution for nonparametric estimation of an underlying distribution function when the observations are measured with error. Depending on the relative size of the measurement error ("noise") to the variability of the underlying distribution of interest ("signal"), it may be impossible to distinguish between very different underlying distributions based on the observed data. Therefore, a nonparametric estimator will not be reliable in the sense that very different looking estimators will be as consistent with the observed data. Surprisingly, the situation is not improved when the sample size is very large (for a given signal-to-noise ratio); the example in Section 2 has an essentially infinite sample size and the application in Section 3 has a sample size of over fifty thousand. We recommend that when using nonparametric methods to estimate the distribution of data measured with error, practitioners perform a sensitivity analysis via simulation to assess how well the nonparametric estimator is able to distinguish different underlying distribution functions, and the extent to which errors in this estimation will affect their conclusions.

Received 10 December 2012

REFERENCES

[1] Billingsley, P. (1986), *Probability and Measure, Second Edition.* New York: Wiley, pp 275–276. MR0830424

[2] Brown, L. D. (2008). In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies, *Annals of Applied Statistics* **2**, 113–152. MR2415597

[3] Carroll, R. J., and Hall, P. (1988). Optimal rates of convergence for deconvolving a density, *Journal of the American Statistical Association* **83**, 1184–1186. MR0997599

[4] Cramer, H. (1936). Uber eine eigenschaft der normalen verteilungsfunktion (in German), *Mathematische Zeitschrift* **41**, 405–414. MR1545629

[5] Dobbin, K. K., Zhao Y., and Simon, R. M. (2008). How large a training set is needed to develop a classifier for microarray data? *Clinical Cancer Research* **14**, 108–114.

[6] Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association* **106**, 1602–1614. MR2896860

[7] Ghosh, M. (1992). Constrained Bayes estimation with applications, *Journal of the American Statistical Association* **87**, 533–540. MR1173817

[8] Groeneboom, P., and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* Basel: Birkhauser Verlag, pp 79–81. MR1180321

[9] Guenther, P. M., Dodd, K. W., Reedy, J., Krebs-Smith S. M., (2006). Most Americans Eat Much Less than Recommended Amounts of Fruits and Vegetables. *Journal of the American Dietetic Association* **106**, 1371–1379.

[10] Kagan, A. M., Linnik, Y. V., and Rao, C. R. (1973), *Characterization Problems in Mathematical Statistics*, New York: Wiley, pp 297–298. MR0346969

[11] Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution, *Journal of the American Statistical Association* **73**, 805–811. MR0521328

[12] Laird, N., and Louis, T. A. (1991). Smoothing the nonparametric estimate of a prior distribution by roughening, *Computational Statistics & Data Analysis* **12**, 27–37. MR1131643

[13] Lenz G., Wright G, Sav S. S. et al. (2008). Stromal Gene Signatures in Large-B-Cell Lymphomas, *New England Journal of Medicine* **359**, 2313–2323.

[14] Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications.* Hayward, CA: Institute of Mathematical Statistics, pp 108–126.

[15] Maloshevskii, S. G. (1968). Sharpness of an estimate of N. A. Sapogov in the stability problem of Cramer's theorem *Theory Of Probility And Its Applications, USSR* **13**, 494–496.

[16] Pilla, R. S., and Lindsay, B. G. (2001). Alternative EM methods for nonparametric finite mixture models, *Biometrika* **88**, 535–550. MR1844850

[17] Robbins, H. (1983). Some thoughts on empirical Bayes estimation, *Annals of Statistics* **11**, 713–723. MR0707923

[18] Sapagov, N. A. (1959). On the independent components of a sum of random variables which is almost normally distributed, *Vestnik Leningrad. Univ.* **14**, 78–105 MR0114232

[19] Scholz, F. W. (1980). Towards a unified definition of maximum likelihood, *Canadian Journal of Statistics* **8**, 193–203. MR0625375

[20] van de Wiel, M. A., and Kim, K. I. (2007). Estimating the false discovery rate using nonparametric deconvolution, *Biometrics* **63**, 806–815 MR2395718

[21] Wright, G., Tan, B., Rosenwald, A., Hurt, E. H., Wiestner, A., Staudt, L. M. (2003). A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9991–9996.

George W. Wright
Biometric Research Branch, MSC 9735
National Cancer Institute
Bethesda, MD 20892
USA
E-mail address: wrightge@mail.nih.gov

Lori E. Dodd
6700A Rockledge Drive #5266
National Institute of Allergy and Infectious Diseases
Bethesda MD 20817
USA
E-mail address: doddl@mail.nih.gov

Edward L. Korn
Biometric Research Branch, MSC 9735
National Cancer Institute
Bethesda, MD 20892
USA
E-mail address: korne@ctep.nci.nih.gov