

Efficient Nonparametric Estimation of a Distribution Function

Reza Modarres*

Department of Statistics
The George Washington University,
Washington DC, 20052, U.S.A.

Abstract

We consider the efficient estimation of a distribution function F under several models based on the Nonparametric likelihood principle. Under the symmetry model, we show that the Nonparametric MLE of F coincides with the symmetrize estimator. Under the auxiliary-sample model, we discuss an estimator based on the total law of probability and show that it coincides with the Nonparametric MLE of F . Assuming quadrant dependence, we show that the estimator has a minimum asymptotic relative efficiency of one with respect to the empirical distribution function. We consider the intersection of the two models and present an efficient hybrid estimator. We show that the hybrid estimator has an asymptotic normal distribution and converges to the Nonparametric MLE of F under the assumption of conditional symmetry. A Monte Carlo simulation assesses the small sample efficiency of the proposed estimators under the Plackett family of bivariate distributions.

1 Introduction

Consider the problem of estimating the distribution function F of a random variable Y with known center θ . There are two cases that are of common interest. The first case concerns estimation of $F(y) = P(Y \leq y)$ for a fixed y ; i.e. estimating a proportion. The second case aims at estimating the entire distribution function. In this article we assume F is unknown and is represented by sample and model information. We discuss estimators that can be used for making pointwise estimates of $F(y)$ for fixed y as well as for every y .

Let y_1, \dots, y_n denote a random sample of size n from F . The empirical distribution function (EDF) $\hat{F}_n(y) = \frac{1}{n} \sum_{i=1}^n I(y_i \leq y)$ is the unique, minimum variance and unbiased estimator of $F(y)$ (Lehman and Casella, 1998). This estimator places probability mass

$\frac{1}{n}$ at each y_i , $i = 1, \dots, n$ and enjoys the property that $n\hat{F}_n(y)$ has a binomial distribution with parameters n and $F(y)$ for each fixed value of y . Thus, $\hat{F}_n(y)$ is unbiased with variance $F(y)(1 - F(y))/n$. EDF is a crude estimator in the sense that it does not take into account any existing model or covariate information.

In order to efficiently estimate $F(y)$ we would like to incorporate existing model or covariate information in the estimation process. Many robust procedures assume the symmetry model, $F(\theta + y) = 1 - F(\theta - y)$ for every y . Assuming that the observations are centered at zero, Shuster (1973) introduced the symmetrize estimator. The symmetrize estimator, $\hat{F}_n^s(y) = \frac{1}{2}(\hat{F}_n(y) + 1 - \hat{F}_n(-y))$, is based on the observation that $\hat{F}_n(y)$ and $\hat{F}_n(-y)$ are antithetic variates (Hammersley and Handscomb, 1964). The estimator mirrors the sample points around zero and places probability mass $\frac{1}{2n}$ at $\pm y_i$, $i = 1, \dots, n$. This results in a minimum of 50% reduction in sample size over $\hat{F}_n(y)$ (see next section). Arcones and Giné (1991) base a powerful bootstrap test of symmetry on $\hat{F}_n^s(y)$. Hinkley (1975) and Taylor (1985) discuss transformations to symmetry in cases where the symmetry models are not justified and Breth (1982) considers the Nonparametric estimation of a symmetric distribution function.

Consider the auxiliary-sample model which assumes access to auxiliary information on a covariate X . The auxiliary information consists of a large sample of m x 's. A random sample of size n , which is usually much smaller than m , bivariate pairs $(x_1, y_1), \dots, (x_n, y_n)$ is collected on (X, Y) from the joint distribution function H . Since m is large one can argue that the marginal distribution K of X is known and equals its empirical distribution function based on m x 's. The information supplied by the auxiliary sample can be used to estimate $F(y)$ more efficiently. It is also of interest to study the case where X provides covariate information to estimate $F(y)$, which is assumed symmetric.

In this paper, we discuss efficient estimation of $F(y) = P(Y \leq y)$ under the above models based on the Nonparametric maximum likelihood principle. We show

*This work has been supported in part by a cooperative agreement with the United States Environmental Protection Agency, Office of Water. E-mail: Reza@gwu.edu

that the proposed estimators are MLE's of $F(y)$ and are asymptotically normal. We obtain the asymptotic variance of these estimators by inverting the Fisher's information matrix. The invariance property of MLE's (Lehman and Casella, 1998) allows us to estimate $t(F)$ using $t(\hat{F})$ for an arbitrary functional t .

In the next section, we show that the Nonparametric MLE for $F(y)$ under the symmetry model coincides with the symmetrize estimator. We discuss efficient estimation of $F(y)$ under the auxiliary-sample model in section 3. Under this model, we show that an estimator based on the total law of probability coincides with the Nonparametric MLE of $F(y)$.

Section 4 considers the intersection of the two models and presents an efficient hybrid estimator. We show that the hybrid estimator is asymptotically normal and converges to the Nonparametric maximum likelihood of $F(y)$ when Y is conditionally symmetry. A Monte Carlo simulation assesses the efficiency of the proposed estimators in the last section.

2 The Nonparametric MLE of $F(y)$ under symmetry

The empirical estimate of the distribution function is a natural estimator of $F(y)$. One can show that the empirical distribution function $\hat{F}_n(y)$ maximizes $L(F) = \prod_{i=1}^n F(\{y_i\})$, where $F(\{y_i\})$ is the probability of the set $\{y_i\}$ and is therefore the Nonparametric maximum likelihood estimate of $F(y)$ (Efron and Tibshirani, 1993, P. 310; Kiefers and Wolfowitz (1956)). Using the same principal one can obtain a Nonparametric estimator of $F(y)$ when the distribution is assumed to be symmetric around zero as follows.

Let $n = \sum_{i=1}^3 n_i$ where n_i represents the observed frequencies corresponding to the probabilities P_i , $i = 1, 2, 3$. Let the sample values be grouped as follows where, for example, $P_1(n_1)$ denotes that there are n_1 observations at or below $-y$ and $P_1 = P(Y \leq -y)$.

$Y \leq -y$	$-y < Y \leq y$	$Y > y$
$P_1 (n_1)$	$P_2 (n_2)$	$P_3 (n_3)$

Let $\hat{F}_n^{mle}(y)$ denote the Nonparametric MLE of $F(y)$ under symmetry. Subject to $\sum_{i=1}^3 P_i = 1$ and $P_1 = P_3$, we have cell configurations $P_1 (n_1), 1 - 2P_1 (n_2)$ and $P_1 (n_3)$. If $y > 0$, then we would like to estimate $F(y) = P_1 + P_2 = 1 - P_1$ and if $y < 0$, we would like to estimate $F(y) = P_1$. The Nonparametric MLE of $F(y)$ is obtained by maximizing the likelihood function $L = P_1^{n_1} (1 - 2P_1)^{n_2} P_1^{n_3}$. The equation for the MLE of

P_1 is $\frac{\partial \log L}{\partial P_1} = \frac{n_1 + n_3}{P_1} - \frac{2n_2}{1 - 2P_1} = 0$. One can verify that $\hat{F}_n^{mle}(y) = \hat{P}_1 = \frac{1}{2n}(n_1 + n_3)$ thus

$$\hat{F}_n^{mle}(y) = \frac{1}{2}(\hat{F}_n(y) + 1 - \hat{F}_n(-y)) = \hat{F}_n^s(y)$$

when $y < 0$ and $\hat{F}_n^{mle}(y) = 1 - \hat{P}_1 = \frac{1}{2n}(n_1 + 2n_2 + n_3) = \hat{F}_n^s(y)$ when $y > 0$. Thus, the Nonparametric MLE of $F(y)$ under symmetry coincides with the symmetrize estimator. That is, $\hat{F}_n^s(y)$ is the MLE of $f(Y)$ under the symmetry model. One can show that $\hat{F}_n^{mle}(y)$ is monotone nondecreasing with $\hat{F}_n^{mle}(-\infty) = 0$ and $\hat{F}_n^{mle}(\infty) = 1$. Thus, it is a genuine distribution function.

Using the large sample properties of MLE's, one can show $\hat{F}_n^{mle}(y)$ is asymptotically normal with mean $F(y)$ and

$$Var(\hat{F}_n^{mle}(y)) = \frac{1}{nI(P_1)} = \frac{1}{2n}P_1(1 - 2P_1)$$

or $Var(\hat{F}_n^{mle}(y)) = \frac{1}{2n}F(-|y|)(1 - 2F(-|y|))$ where $P_1 = F(y)$ if $y < 0$ and $P_1 = 1 - F(y) = F(-y)$ if $y > 0$ and $I(P_1) = -E_{P_1}(\frac{\partial^2}{\partial P_1^2} \log L)$ is the Fisher information of P_1 . In order to compare the estimators let $\rho(\hat{F}_n(y), \hat{F}_n^{mle}(y)) = \frac{var(\hat{F}_n(y))}{var(\hat{F}_n^{mle}(y))}$. One can show $\rho(\hat{F}_n(y), \hat{F}_n^{mle}(y)) = \frac{2(1-F(-|y|))}{1-2F(-|y|)} \geq 2$ when Y is symmetric. One can plot a graph of the efficiency, relative to EDF, for a given choice of F . The graph is similar to a double-exponential curve with asymptotic tails at 2.0. The efficiency is greatest near the center where it becomes infinite.

3 The Nonparametric MLE of $F(y)$ with auxiliary samples

Several authors have used auxiliary information to improve the estimation of a distribution function. Under a finite population model, Rao, Kovar and Mantel (1990) proposed design-based estimators of the distribution function. Kuk (1993) notes that the proposed estimators depend on an implicit linearity assumption between X and Y , and proposes an interesting method by combining the known distribution of the auxiliary variable with a kernel estimate of the conditional distribution function of Y given X . Viewed in a finite population model, the estimator we will discuss is based on dichotomizing the Y values and does not depend on the existence of a linear relationship between X and Y . In fact, it will be shown that quadrant dependence is sufficient to ensure a minimum asymptotic relative efficiency

of one with respect to the empirical distribution function. In an interesting application, Rothery (1982) discusses the use of auxiliary information for estimating the power of a non-parametric test of hypothesis. Ocaña and Vegas (1995) show that the maximum likelihood estimator of the power function coincides with the regression-adjusted estimator. Koshevink (1994) discusses the Non-parametric estimation of a distribution function in stratified data under a biased sampling model. Hesterberg and Nelson (1998) explore the use of stratified sampling to estimate the distribution and quantile functions. In particular, their discrete approximation yields the non-parametric MLE of $F(y)$. To incorporate the auxiliary information we will condition on the auxiliary-sample and apply the law of total probability as follows.

We assume that the m auxiliary-sample observations fall into $r+2$ classes labeled C_i , $i = 0, \dots, r+1$ for $r \geq 0$. The basic idea is post stratification of the sample in order to increase the efficiency of the estimator (Cochran, 1977). We will classify Y as $Y \leq y$ or $Y > y$ according to the value of auxiliary X and form a linear combination of the conditional distribution functions. Let the n observed pairs (x_i, y_i) divide the (X, Y) plane into $2(r+2)$ regions as follows.

X	$Y \leq y$	$Y > y$
$C_0 = \{x x \leq x_0\}$	$P_0 (n_0)$	$P_{r+2} (n_{r+2})$
.....
$C_i = \{x x_{i-1} < x \leq x_i\}$	$P_i (n_i)$	$P_{i+r+2} (n_{i+r+2})$
.....
$C_{r+1} = \{x x > x_r\}$	$P_{r+1} (n_{r+1})$	$P_{2r+3} (n_{2r+3})$

Let $n = \sum_{i=0}^{2r+3} n_i$ where n_i denote the observed frequencies with the corresponding probabilities $P_i = P(Y \leq y, X \in C_i)$ and $P_{i+r+2} = P(Y > y, X \in C_i)$, $i = 0, \dots, r+1$. Conditioning on the auxiliary sample is a natural approach when estimating $F(y)$. By the law of total probability we have $F(y) = \sum_{i=0}^{r+1} P_i = \sum_{i=0}^{r+1} F(y|X \in C_i)P(X \in C_i)$. Now, if $F(y|X \in C_i)$, $i = 0, \dots, r+1$ are estimated from the joint sample (x_i, y_i) , $i = 1, \dots, n$, one has

$$\hat{F}^{aux}(y) = \sum_{i=0}^{r+1} \frac{n_i}{n_i + n_{i+r+2}} P(X \in C_i).$$

The auxiliary information supplied through $P(X \in C_i)$ is used to reweigh the estimated conditional probabilities. Thus, $\hat{F}^{aux}(y)$ is a weighted average of $r+2$ estimated conditional distribution functions. One may show that $\hat{F}^{aux}(y)$ coincides with the Nonparametric MLE of $F(y)$ as follows.

The Nonparametric MLE of $F(y) = \sum_{i=0}^{r+1} P_i$ is obtained by maximizing the likelihood function $L = \prod_{i=0}^{2r+3} P_i^{n_i}$ subject to $\sum_{i=0}^{2r+3} P_i = 1$ and known $P(X \in C_i) = P_i + P_{i+r+2}$. The i th likelihood equation is $\frac{\partial \log L}{\partial P_i} = \frac{n_i}{P_i} - \frac{n_{i+r+2}}{P(X \in C_i) - P_i} = 0$, with solution $\hat{P}_i = \frac{n_i}{n_i + n_{i+r+2}} P(X \in C_i)$. Thus, $\hat{F}^{aux}(y) = \sum_{i=0}^{r+1} \hat{P}_i = \hat{F}^{mle}(y)$ and the two estimators are the same.

In order to show $\hat{F}^{mle}(y)$ is a genuine distribution function for fixed x 's we note that $\hat{F}_n^{aux}(+\infty) = \sum_{i=0}^{2r+3} P_i = 1$ because $P_{i+r+2} = 0$ and $\hat{F}_n^{aux}(-\infty) = 0$ because $P_i = 0$, $i = 0, \dots, r+1$. To show that $\hat{F}_n^{aux}(y)$ is monotone nondecreasing we need to show $y_1 \leq y_2$ implies $\hat{F}_n^{aux}(y_1) \leq \hat{F}_n^{aux}(y_2)$. Consider two tables with the observed frequencies, n_i and n_i^* , $i = 0, \dots, 2r+3$. The tables are based on the same sample of size n but are evaluated at y_1 and y_2 where $y_1 \leq y_2$. We have $n_i \leq n_i^*$, $i = 0, \dots, r+1$ when $y_1 \leq y_2$. $\hat{F}_n^{aux}(y_1) \leq \hat{F}_n^{aux}(y_2)$ follows from the definition of $\hat{F}_n^{aux}(y)$ for fixed x 's.

Using large sample properties of MLE's, one can show that $\hat{F}_n^{aux}(y)$ is asymptotically normal with mean $F(y)$. The asymptotic variance is obtained from the inverse of Fisher's information matrix of \hat{P}_i 's

$$Var(\hat{F}_n^{aux}(y)) = \frac{1}{n} \sum_{i=0}^{r+1} \frac{P_i(P(X \in C_i) - P_i)}{P(X \in C_i)}.$$

We would like to study conditions under which $Var(\hat{F}_n^{aux}(y)) \leq Var(\hat{F}_n(y))$. Note that

$$Var(\hat{F}_n^{aux}(y)) - Var(\hat{F}_n(y)) = \frac{1}{n} (F^2(y) - \sum_{i=0}^{r+1} \frac{P_i^2}{P(X \in C_i)}).$$

If we assume $P_i = P(X \in C_i, Y \leq y) \geq P(Y \leq y)P(X \in C_i)$ for $i = 0, \dots, r+1$, then it can be verified that $Var(\hat{F}_n^{aux}(y)) \leq Var(\hat{F}_n(y))$. The set of conditions $P_i \geq P(Y \leq y)P(X \in C_i)$ lead to

$$\sum_{j=0}^i P_j = \sum_{j=0}^i P(X \in C_j, Y \leq y) = H(x_i, y)$$

where $H(x_i, y) \geq P(Y \leq y) \sum_{i=0}^i P(X \in C_i) = F(y)K(x_i)$. Conditions $H(x_i, y) \geq F(y)K(x_i)$ for $i = 0, \dots, r+1$ are true when X and Y are positively quadrant dependent (Lehmann (1966)); i.e. $H(x, y) = P(X \leq x, Y \leq y) \geq P(Y \leq y)P(X \leq x) = F(y)K(x)$ for all $(x, y) \in R^2$. One can show that when X and Y are positively quadrant dependent $Cov(X, Y) \geq 0$ (Nelsen, 1999, p. 153). It can be verified that $\rho(\hat{F}_n(y), \hat{F}_n^{aux}(y)) = var(\hat{F}_n(y))/var(\hat{F}_n^{aux}(y)) \geq 1$ whenever $F^2(y) \leq \sum_{i=0}^{r+1} \frac{P_i^2}{P(X \in C_i)}$. In particular, we note that $\rho(\hat{F}_n(y), \hat{F}_n^{aux}(y)) = 1$ when X and Y are independent.

To estimate $F(y|X \in C_i)$ with precision, $n_i + n_{i+r+2}$ must be large. To that end one can increase n or decrease the number of classes r . Since the number of pairs (x_i, y_i) is usually small it is desirable to reduce the number of classes to two or three. To form two classes we propose to select x_s from the m x 's and form the following 2×2 table.

	$Y \leq y$	$Y > y$
$X \leq x_s$	$P_0 (n_0)$	$P_2 (n_2)$
$X > x_s$	$P_1 (n_1)$	$P_3 (n_3)$

We would like to select x_s such that X and Y become most concordant. We consider the boundary distributions

$H_1(x, y) = \min(F(y), K(x))$ and $H_{-1}(x, y) = \max(F(y) + K(x) - 1, 0)$ where $H_{-1}(x, y) \leq H(x, y) \leq H_1(x, y)$ (Mardia, 1967, P. 30) and select x_s to maximize the upper bound $H_1(x, y) = \frac{1}{2}(F(y) + K(x) - |F(y) - K(x)|)$. It can be seen that $H_1(x, y)$ degenerates on the non-decreasing curve $F(y) = K(x)$. Thus, $x = K^{-1}(F(y))$ will maximize the upper bound and a sensible selection strategy is to select x such that $\delta = |F(y) - K(x)|$ is minimum. In particular, we select the smallest x_s such that $x_s = \hat{K}^{-1}(\hat{F}_n(y))$ where \hat{K}^{-1} is the quantile function of X based on m auxiliary-sample observations. This is, in fact, the $m\hat{F}_n(y)$ th smallest ordered value of the m x 's. Since the EDF is a crude estimate of $F(y)$ one may consider the selection of more than one x_s . A simple procedure is to consider two values of y and find the corresponding x_v and x_s where $v < s$ and use the results for the case $r = 1$.

Consider the conditions of positive quadrant dependence in the 2×2 case where $r = 0$. The conditions $P_i = P(x \in C_i, y \leq y) \geq P(x \in C_i)P(y \leq y)$, for $i = 0, 1$ lead to $\pi = \frac{P_0 P_3}{P_2 P_1} \geq 1$ where π is the odds ratio. One can show that, in the 2×2 case, (X, Y) are positively quadrant dependent if and only if $\pi \geq 1$ (Nelsen, 1999, P. 153; Lehmann, 1986, P. 176). We will consider the 2×2 case further in section 5, where we discuss a Monte Carlo study.

4 A hybrid estimator for the models at the intersection

In this section, we will use auxiliary as well as model information to estimate $F(y)$. We assume that $(x_i, y_i), i = 1, \dots, n$ is a random sample from $H(X, Y)$ and an auxiliary sample of m X 's is available in grouped form with $r+2$ classes. We also assume that m is large relative to n so that $P(X \in C_i), i = 0, \dots, r+1$ is known and Y is marginally symmetric about a known center θ , taken to be zero. Let the $n = \sum_{i=0}^{3r+6} n_i$ observed pairs divide the (X, Y) plane into $3(r+2)$ regions as follows.

X	$Y \leq -y$	$-y < Y \leq y$	$Y > y$
C_0	$P_0(n_0)$	$P_{r+2}(n_{r+2})$	$P_{2r+4}(n_{2r+4})$
...
C_i	$P_i(n_i)$	$P_{i+r+2}(n_{i+r+2})$	$P_{i+2r+4}(n_{i+2r+4})$
...
C_{r+1}	$P_{r+1}(n_{r+1})$	$P_{2r+3}(n_{2r+3})$	$P_{3r+5}(n_{3r+5})$

In order to obtain the Nonparametric MLE of $F(y)$ one needs to maximize the likelihood function $L = \prod_{i=0}^{3r+5} P_i^{n_i}$ subject to $\sum_{i=0}^{3r+5} P_i = 1, P_i + P_{i+r+2} + P_{i+2r+4} = P(X \in C_i), i = 0, \dots, r+1$ and $\sum_{i=0}^{r+1} P_i = \sum_{i=2r+4}^{3r+5} P_i$. However, the likelihood function does not have a solution. One may instead combine the auxiliary-sample estimator and the symmetrize estimator of $F(y)$ to produce a hybrid estimator as follows.

When $y > 0$ we would like to estimate $F(y) = \sum_{i=0}^{2r+3} P_i$. Maximizing the likelihood function subject to the first two

conditions results in $\hat{P}_i = \frac{n_i}{n_i + n_{i+r+2} + n_{i+2r+4}} P(X \in C_i)$. The auxiliary-sample estimator of $F(y)$ is $\hat{F}^{aux}(y) = \sum_{i=0}^{2r+3} \hat{P}_i$ and the symmetrize estimator of $F(y)$ is $\hat{F}_n^s(y) = \frac{1}{2}(F_n(y) + 1 - \hat{F}_n(-y))$. Since Y is symmetric around zero, a more efficient estimator is formed if one symmetrize an improved estimate of $F(y)$. Thus, the hybrid estimator is obtained from $\hat{F}^h(y) = \frac{1}{2}(\hat{F}^{aux}(y) + 1 - \hat{F}^{aux}(-y))$ or

$$\hat{F}^h(y) = \frac{1}{2} + \sum_{i=0}^{r+1} \frac{n_{i+r+2}}{2(n_i + n_{i+r+2} + n_{i+2r+4})} P(X \in C_i).$$

When $y < 0$ one estimates $F(y) = \sum_{i=0}^{r+1} P_i$ by

$$\hat{F}^h(y) = \frac{1}{2} - \sum_{i=0}^{r+1} \frac{n_{i+r+2}}{2(n_i + n_{i+r+2} + n_{i+2r+4})} P(X \in C_i).$$

Note that $\hat{F}^h(y)$ is bounded below or above at $\frac{1}{2}$ depending on whether $y \leq 0$ or $y > 0$, respectively. It equals $\frac{1}{2}$ when $n_{i+r+2} = 0$ for $i = 0, \dots, r+1$; thus, suggesting zero is the center of symmetry. To gain insight in the behavior of the hybrid estimator note that $P(Y \leq -y|X \in C_i) = P(Y > y|X \in C_i)$ for all i and y implies that Y is symmetric around zero. We consider the conditional distribution of Y given $X \in C_i$ and assume that it is symmetric about $E(Y|X \in C_i)$. If one further assumes that $E(Y|X \in C_i)$ is known and set to zero, one can show the hybrid estimator coincides with the Nonparametric MLE of $F(y)$ as follows.

When $y > 0$ an estimate for $F(y) = \sum_{i=0}^{2r+3} P_i$ is obtained by maximizing the likelihood function $L = \prod_{i=0}^{3r+5} P_i^{n_i}$ subject to $\sum_{i=0}^{3r+5} P_i = 1, P_i + P_{i+r+2} + P_{i+2r+4} = P(X \in C_i)$, and $P_i = P_{i+2r+4}, i = 0, \dots, r+1$. The resulting estimator of $F(y) = 1 - F(-y) = 1 - \sum_{i=0}^{r+1} P_i$ is

$$\hat{F}^{mle}(y) = 1 - \sum_{i=0}^{r+1} \frac{n_i + n_{i+2r+4}}{2(n_i + n_{i+r+2} + n_{i+2r+4})} P(X \in C_i).$$

when $y < 0$, we have $F(-y) = \sum_{i=0}^{r+1} P_i$ which is estimated as

$$\hat{F}^{mle}(y) = \sum_{i=0}^{r+1} \frac{n_i + n_{i+2r+4}}{2(n_i + n_{i+r+2} + n_{i+2r+4})} P(X \in C_i).$$

One can verify that $F^h(y) = \hat{F}^{mle}(y)$ for all y . However, in most cases we do not know $E(Y|X \in C_i)$. If this is estimated by an unbiased and consistent estimator, then the above argument holds asymptotically; i.e. as $\min(n_i)$ approaches ∞ . Thus, the hybrid estimator obtained by combining the symmetrize estimator and the auxiliary-sample estimator is asymptotically the same as the Nonparametric MLE, assuming that Y is conditionally symmetric.

Using the large sample properties of MLE's, one can show $\hat{F}^h(y)$ is asymptotically normal with mean $F(y)$ and variance

$$Var(\hat{F}^h(y)) = \frac{1}{2n} \sum_{i=0}^{r+1} \frac{P_i(P(X \in C_i) - 2P_i)}{P(X \in C_i)}$$

obtained from the inverse of Fisher's variance matrix of P_i 's. To compare the hybrid and the EDF estimators define $\rho(\hat{F}_n(y), \hat{F}^h(y)) = \text{Var}(\hat{F}_n(y))/\text{Var}(\hat{F}^h(y))$. One can show $\rho(\hat{F}_n(y), \hat{F}^h(y)) = 2(F(y) - F^2(y))/(F(-|y|) - 2\sum_{i=0}^{r+1} P_i^2/P(x \in C_i))$. When X and Y are independent and Y is symmetric, one can show that the hybrid estimator reduces to the symmetrize estimator and verify

$$\rho(\hat{F}_n(y), \hat{F}^h(y)) = \rho(\hat{F}_n(y), \hat{F}^s(y)).$$

Thus, the hybrid estimator is robust with respect to the independence assumption, but not with respect to the symmetry assumption. In fact, our simulation results show that $\rho(\hat{F}_n(y), \hat{F}^s(y))$ and $\rho(\hat{F}_n(y), \hat{F}^h(y))$ are less than unity when Y is not symmetric. Note that

$$\text{Var}(\hat{F}^h(y)) - \text{Var}(\hat{F}_n^s(y)) = \frac{1}{n}(F^2(-|y|) - \sum_{i=0}^{r+1} \frac{P_i^2}{P(X \in C_i)}).$$

One can show $\text{Var}(\hat{F}^h(y)) \leq \text{Var}(\hat{F}_n^s(y))$ when $H(x_i, y) \geq F(y)K(x_i)$, which are true when X and Y are positively quadrant dependent. In fact, $\text{Var}(\hat{F}^h(y)) = \text{Var}(\hat{F}_n^s(y))$ when X and Y are independent.

5 Monte Carlo Simulation

In this section, some simulation results are summarized for the purpose of comparing the estimators and their efficiency. Plackett's family of distributions is used to represent the joint distribution of the variables X and Y . Plackett's family of bivariate distributions is composed of all distribution functions $H(x, y)$ which satisfy

$$\pi = \frac{H(x, y)(1 - K(x) - F(y) + H(x, y))}{(K(x) - H(x, y))(F(y) - H(x, y))}$$

where K and F are arbitrary marginal distribution functions and odds ratio $\pi \in (0, \infty)$. Thus, we can model a variety of marginal with a full range of dependence. It can be shown (Nelson, 1999, p. 153) that $\pi > 1$ when X and Y are positively quadrant dependent and $\pi = 1$ when they are independent.

A Monte Carlo simulation based on 2500 replications was performed to estimate $P(Y \leq y)$, where y was obtained from $F(y) = p$, for $p = 0.05, \dots, 0.95$ with increment of 0.05. $F(y) = 0.50$ was excluded as the relative efficiency becomes infinite at the center for several of the estimators. We generated $n = 100$ bivariate samples from a Plackett distribution with $\pi = 0$ and $\pi = 100$ and marginal distributions: *i*) Normal with mean zero and unit variance, *ii*) Uniform in the interval $(0, 1)$ and *iii*) Exponential with mean 1. The above nine bivariate distributions have specified marginal and dependence structure measured by the odds-ratio (when $r=0$). The distribution $K(x)$ is estimated from an auxiliary sample of size $m = 1000$ when $\pi = 0$ and $m = 2000$ when $\pi = 100$. The smallest x_s that satisfies the selection rule $\hat{k}(x_s) > \hat{F}(y)$ was used to form the auxiliary-sample estimators. In particular, for the 2×2 case, we used $x_s = \hat{k}^{-1}(n\hat{F}_n(y))$ and for

the 2×3 and 3×3 cases we used $x_s = \hat{k}^{-1}(n\hat{F}_n^s(y))$. For 3×2 and 3×3 cases, x_s was set to the sth order statistic and v was set to $s - 1$. To ensure three classes, when $s = 1$ or $s = n$ we used $s = 2$ or $s = n - 1$ with $v = 1$ or $v = n$, respectively.

The following estimators were compared in terms of bias and relative efficiency: EDF \hat{F}_n , symmetrize \hat{F}_n^s , the auxiliary-sample estimator for a 2×2 table $\hat{F}_{2 \times 2}^{aux}$, the hybrid estimator $\hat{F}_{2 \times 3}^h$, the auxiliary-sample estimator $\hat{F}_{3 \times 2}^{aux}$ and the hybrid estimator $\hat{F}_{3 \times 3}^h$. The bias and the relative efficiencies with respect to the EDF estimator for the nine bivariate distributions were obtained. Results for the case where X has an exponential distribution are similar to those where X has a uniform or normal distribution. We only report the results for the case where X has an exponential distribution and Y has a normal, uniform and exponential distribution. Figures 1-3 show plots of bias against $p = F(y)$ and Figures 4-6 show plots of the relative efficiency against p for $\pi = 1$. The corresponding graphs for $\pi = 100$ appear in Figures 7-12.

Figures 1-2 show that the bias of all the estimators lie in the range $(-0.004, 0.003)$ when Y is symmetric. Figure 3 conveys that $\hat{F}_{2 \times 2}^{aux}$, $\hat{F}_{2 \times 3}^h$, and \hat{F}_n^s have little bias whereas the symmetry-based estimators $\hat{F}_{3 \times 3}^h$, \hat{F}_n^s , and $\hat{F}_{2 \times 3}^h$ show substantial bias in the range $(0.06, -0.14)$. The reason for the bias is the fact that Y is asymmetric in this case. Figures 4-5 show the finite sample relative efficiency for \hat{F}_n^s and $\hat{F}_{2 \times 3}^h$ are identical and followed by a graph of $\hat{F}_{3 \times 3}^h$. In this case $\theta = 1$ implies that the auxiliary sample does not provide more information to the hybrid estimator. The graph of relative efficiency is similar to a double-exponential curve with asymptotic tails at 2.0. $\hat{F}_{2 \times 2}^{aux}$ and $\hat{F}_{3 \times 2}^{aux}$ behave similar to the EDF. Figure 6 show that the small sample relative efficiency falls below 1.0 for symmetry-based estimators when Y is asymmetric.

Figures 7-8 show that the bias of all estimators lie in the range $(-0.008, 0.006)$ when Y is symmetric. $\hat{F}_{2 \times 2}^{aux}$ and $\hat{F}_{3 \times 2}^{aux}$ behave similarly and show more bias than others. Figure 9 is similar to Figure 3 and shows much more bias for symmetry-based estimators when Y is asymmetric. Figures 10-11 show similar behavior for symmetry-based estimators. If one compares Figures 10-11 with Figures 4-5 one observes that symmetry provides as much information as an auxiliary sample of size m when $\theta = 100$. Figures 10-12 show that $\hat{F}_{2 \times 2}^{aux}$, and $\hat{F}_{3 \times 2}^{aux}$ have a finite sample relative efficiency of at least 1 and at most 3.0. Figure 12 shows relative efficiencies which are below 1.0 for all symmetry-based estimators.

References

- [1] Arcones, M. A. and Giné, E. (1991). Some Bootstrap Tests of Symmetry for Univariate Continuous Distributions. *Ann. Statist.* 19, 1496-1511.
- [2] Breth M., (1982). Nonparametric Estimation for a Symmetric Distribution. *Biometrika* 69, 625-634.
- [3] Cochran, W. G. (1977). *Sampling techniques*, Third edi-

tion, John Wiley & Sons, New York.

[4] Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

[5] Hammersley J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods*, London: Methuen and Co., Ltd.

[6] Hesterberg, T. C. and Nelson, B. (1998). Control Variates for Probability and Quantile Estimation. *Management Science*, Vol. 44, No. 9, 1296-1312.

[7] Hinkley, D. V. (1975). On Power Transformations to Symmetry. *Biometrika* 62, 01-111.

[8] Kiefer, j. and Wolfowitz, J. (1956). Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Nuisance Parameters. *Annals of Mathematical Statistics*, 27, 887-906.

[9] Koshevnik, Y. (1994). Nonparametric CDF Estimation from Stratified Data. *Proceedings of the 26th Symposium on the Interface: Computationally Intensive Statistical Methods*. John Sall and Ann Lehman, Editors. 459-463.

[10] Kuk, Anthony Y. C. (1993). A Kernel Method for Estimating Finite Population Distribution Functions Using Auxiliary Information. *Biometrika*, 80,2, 385-392.

[11] Lehman, E. L. and Casella G. (1998). *Theory of Point Estimation*. Second edition, John Wiley & Sons, New York.

[12] Mardia, K. V. (1967). *Families of Bivariate Distributions*. Charles Griffin and Co., London.

[13] Nelsen, Roger B. (1999). *An Introduction to Copulas*. Springer.

[14] Ocaña, J. and Vegas, E. (1995). Variance Reduction for Bernoulli Response Variables in Simulation. *Computational Statistics and Data Analysis*, 19, 631-640.

[15] Rao, J. N. K., Kovar, J. G. and Mantel, H. J. (1990). On Estimating Distribution Functions and Quantiles from Survey Data Using Auxiliary Information. *Biometrika*, 77, 2, 565-375.

[16] Rothery, P. (1982). The Use of the Control Variates in Monte Carlo Estimation of Power. *Applied Statistics*, 31, 125-129.

[17] Shuster, E. F. (1973). On the Goodness-of-Fit Problem for Continuous Symmetric Distributions. *J. Am. Statist. Assoc.* 68, 713-5. Corrigenda (1974). *J. Am. Statist. Assoc.* 69, 288.

[18] Taylor, M. G. (1985). Power Transformations to Symmetry. *Biometrika* 72, 145-152.

Figure 1: Bias for $(X, Y) = (\text{Exp.}, \text{Normal})$ with $\theta = 1$ and $m = 1000$

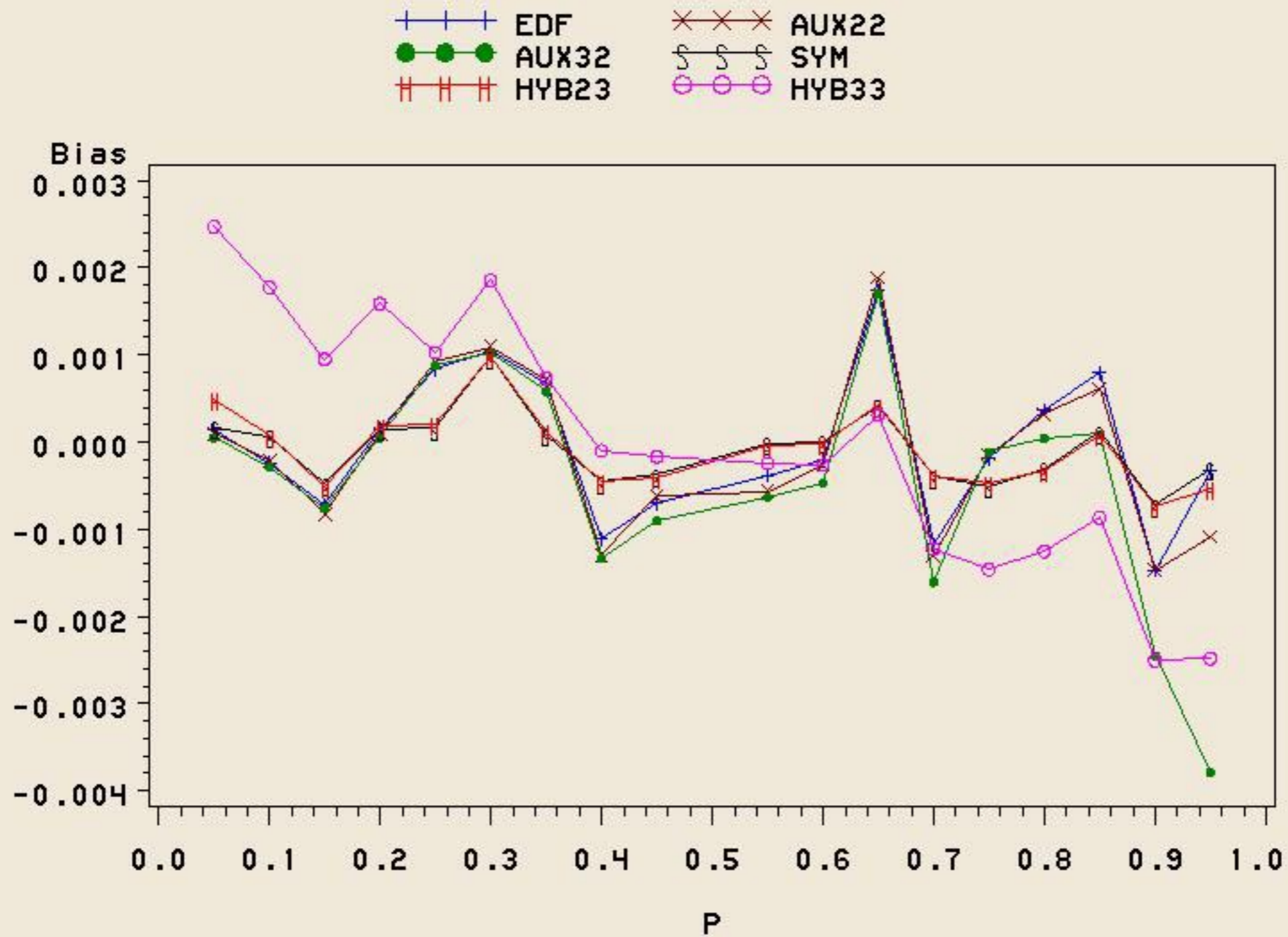


Figure 2: Bias for $(X,Y)=(Exp., Uniform)$ with $\theta=1$ and $m=1000$

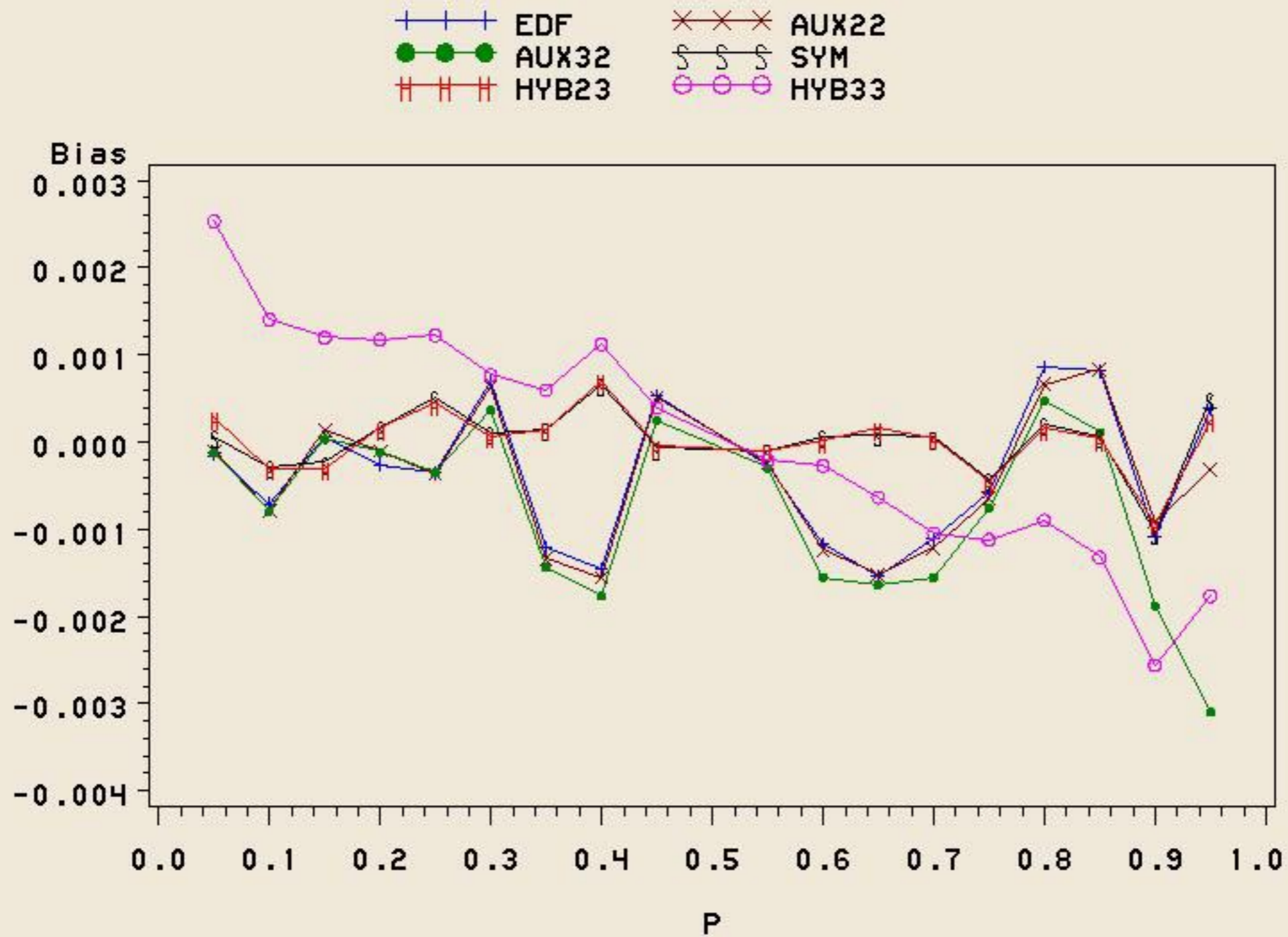


Figure 3: Bias for $(X, Y) = (\text{Exp.}, \text{Exp.})$ with $\theta = 1$ and $m = 1000$

+ + +	EDF	× × ×	AUX22
● ● ●	AUX32	§ § §	SYM
H H H	HYB23	○ ○ ○	HYB33

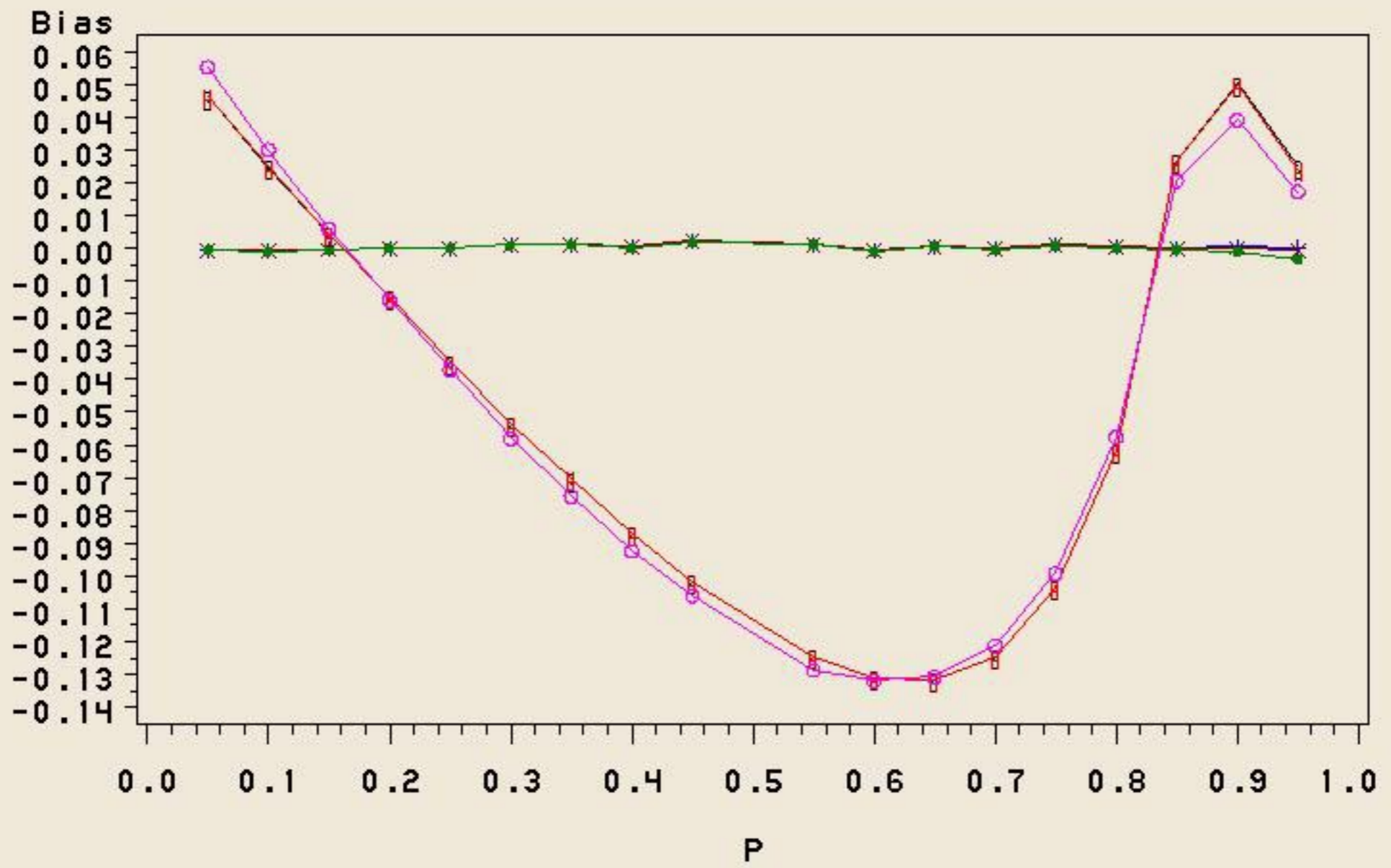


Figure 4: Relative Efficiency for $(X, Y) = (\text{Exp.}, \text{Normal})$ with $\theta = 1$ and $m = 1000$

××× AUX22 ●●● AUX32
— — — SYM ††† HYB23
○ ○ ○ HYB33

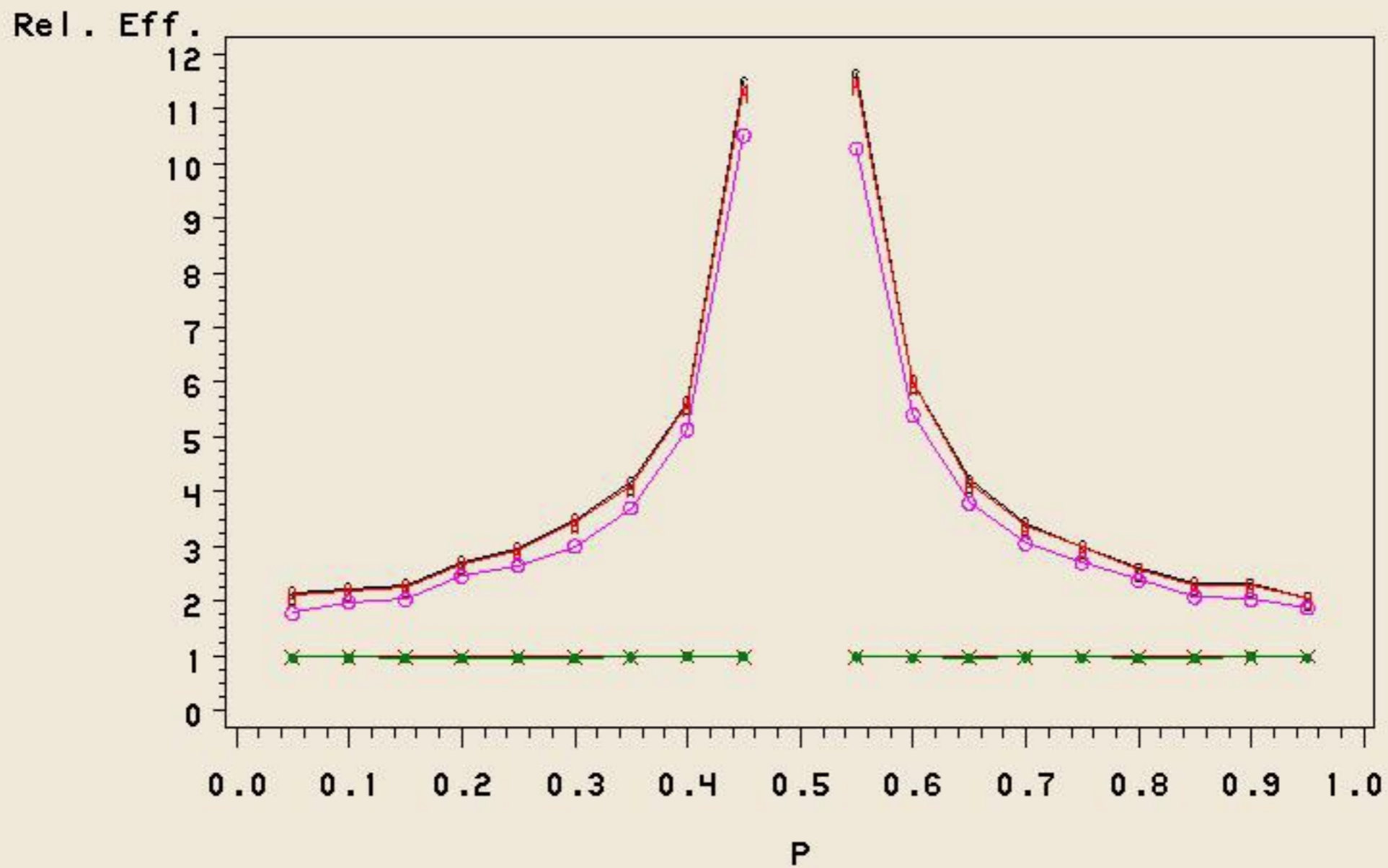


Figure 5: Relative efficiency for $(X, Y) = (\text{Exp.}, \text{Uniform})$ with $\theta = 1$ and $m = 1000$

××× AUX22 ●●● AUX32
—S—S—S SYM H—H—H HYB23
○—○—○ HYB33

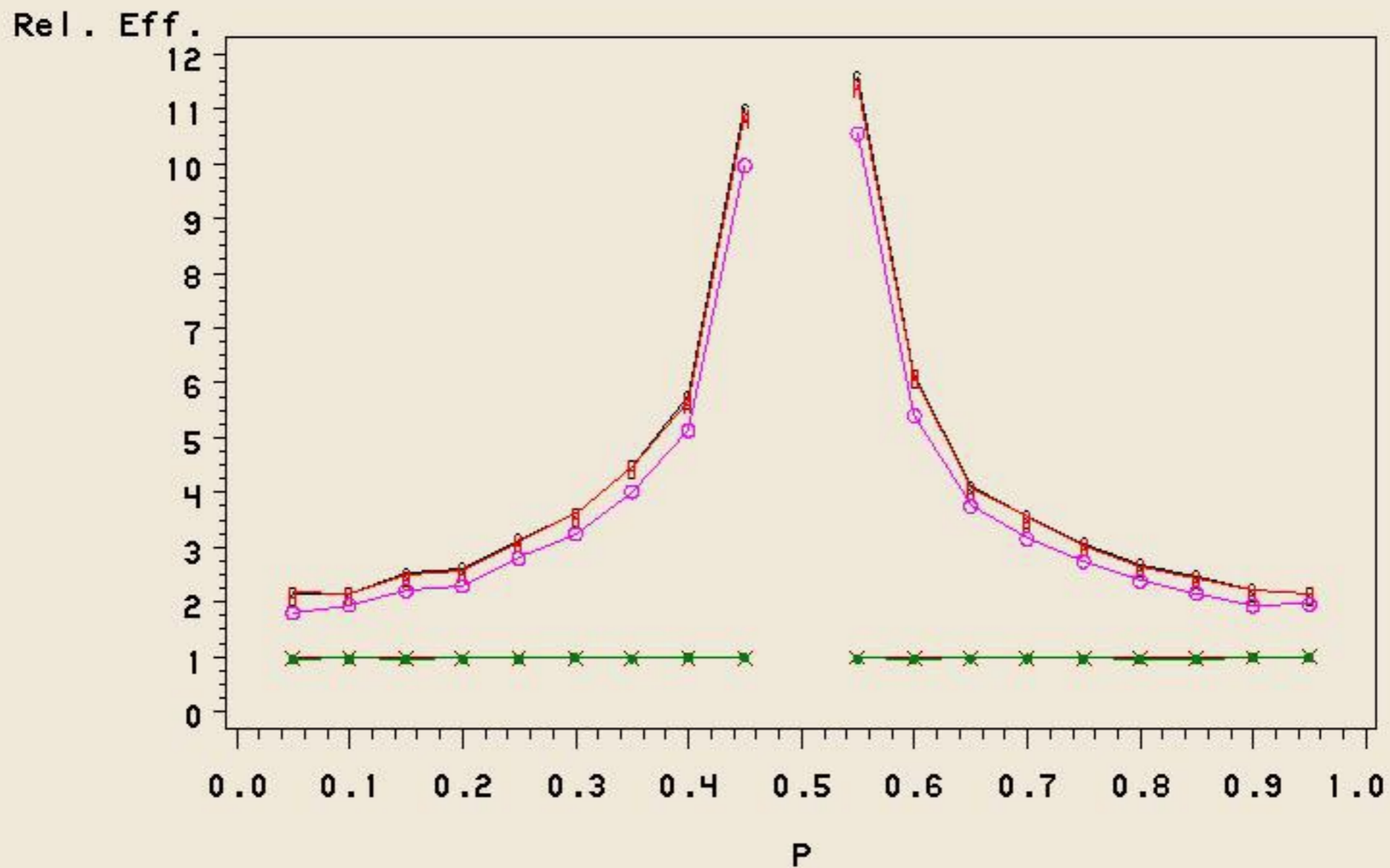


Figure 6: Relative Efficiency for $(X, Y) = (Exp., Exp.)$ with $\theta = 1$ and $m = 1000$

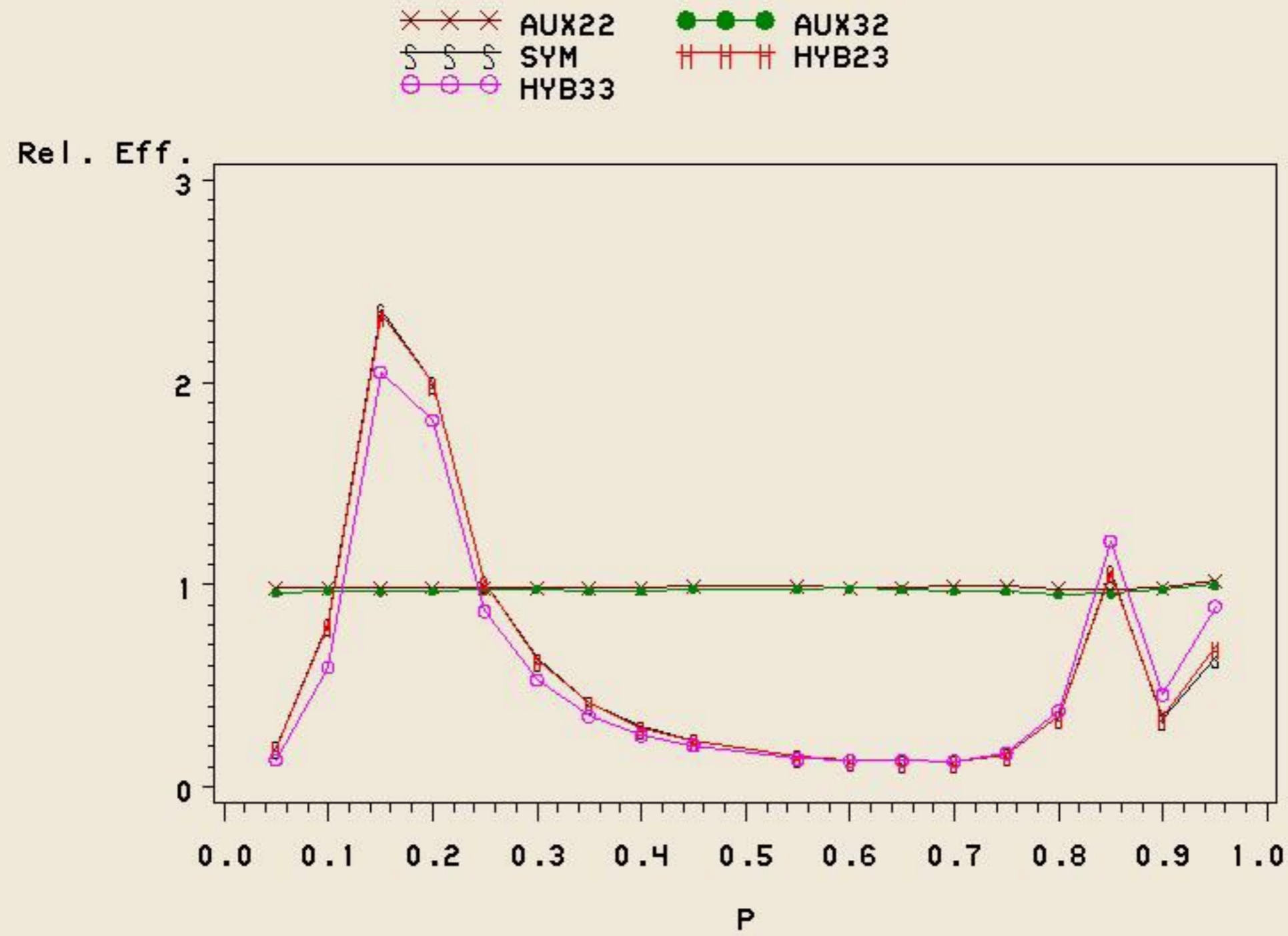


Figure 7: Bias for $(X, Y) = (\text{Exp.}, \text{Normal})$ with $\theta = 100$ and $m = 2000$

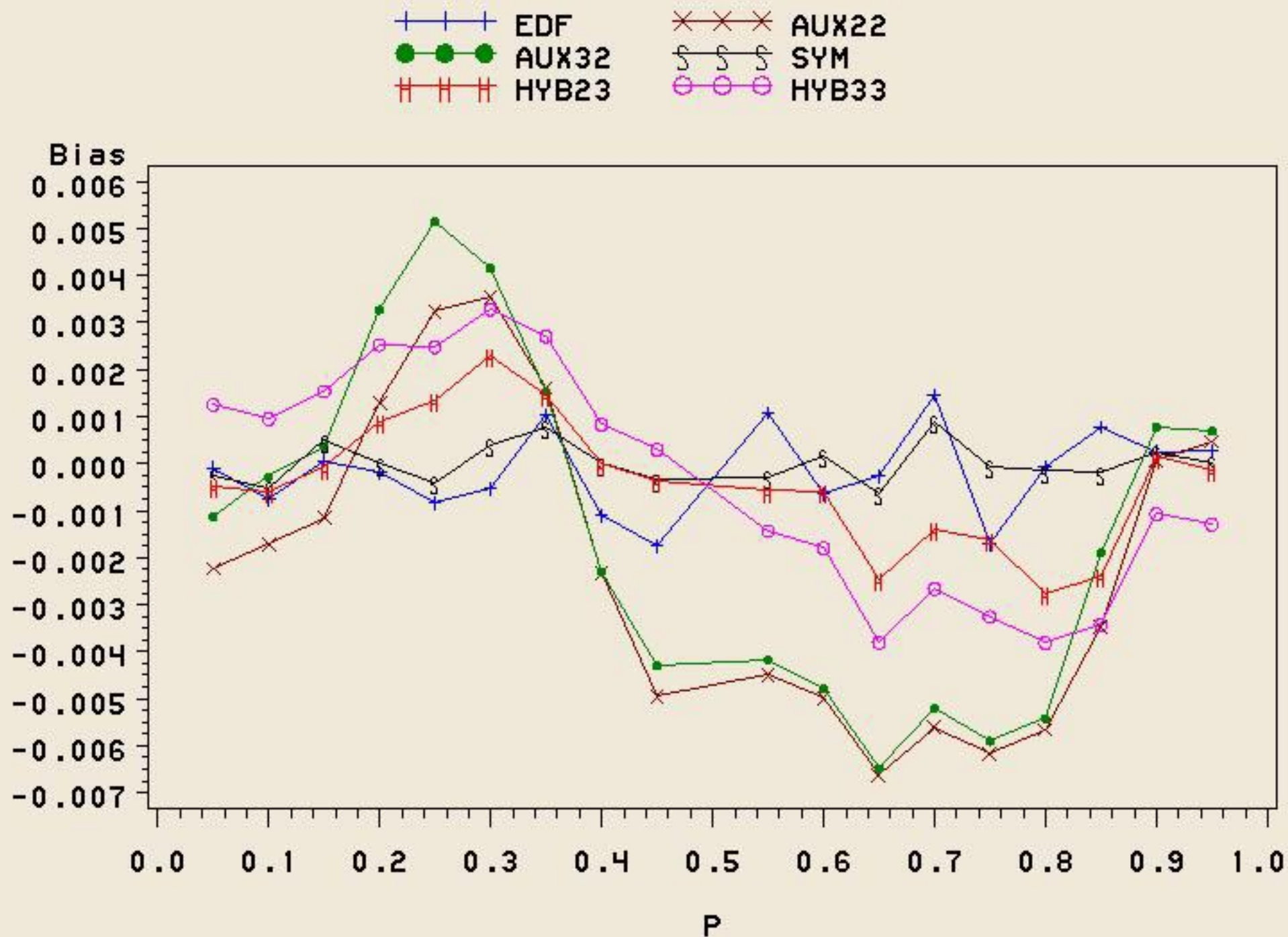


Figure 8: Bias for $(X, Y) = (\text{Exp.}, \text{Uniform})$ with $\theta = 100$ and $m = 2000$

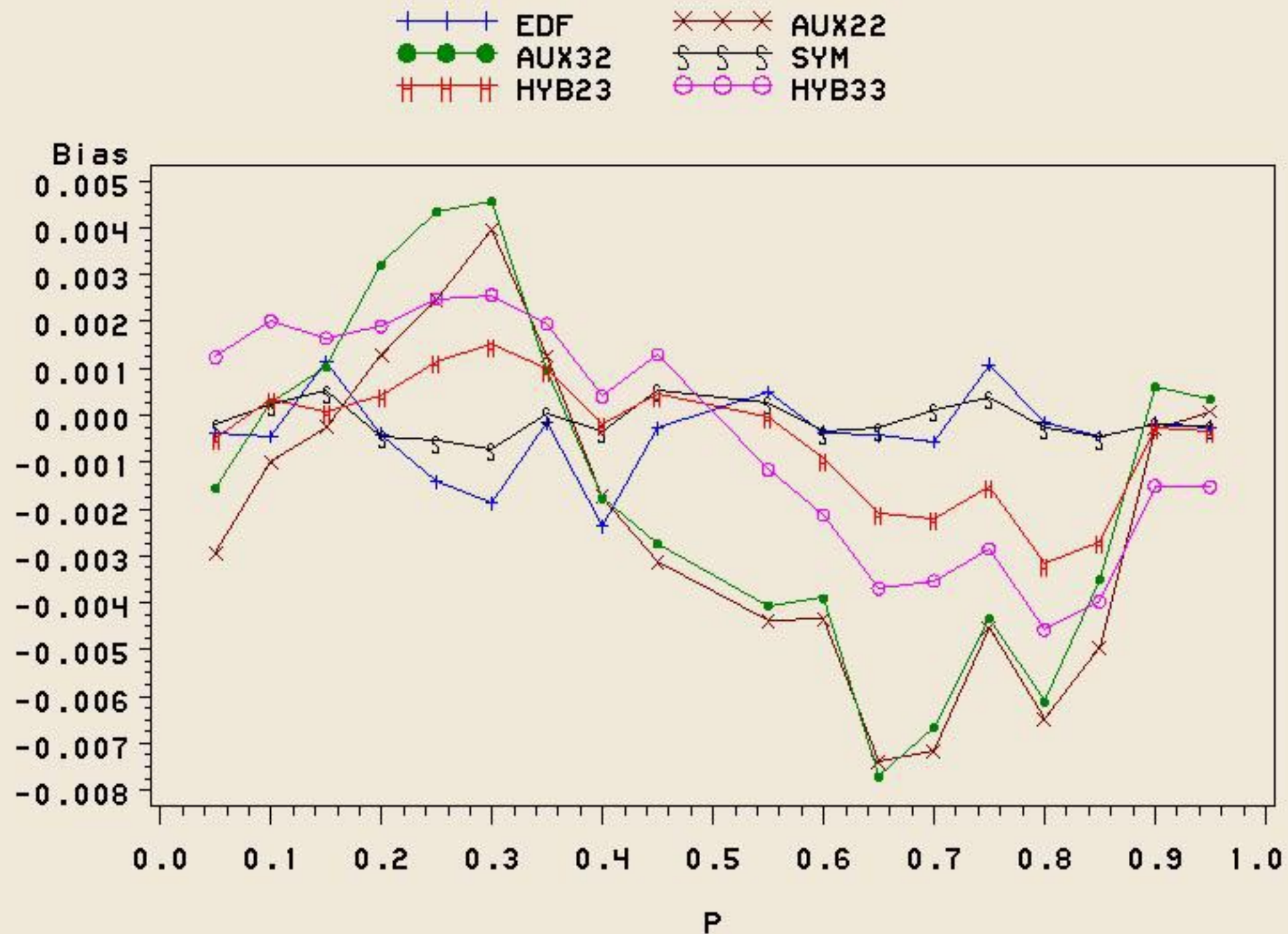


Figure 9: Bias for $(X, Y) = (\text{Exp.}, \text{Exp.})$ with $\theta = 100$ and $m = 2000$

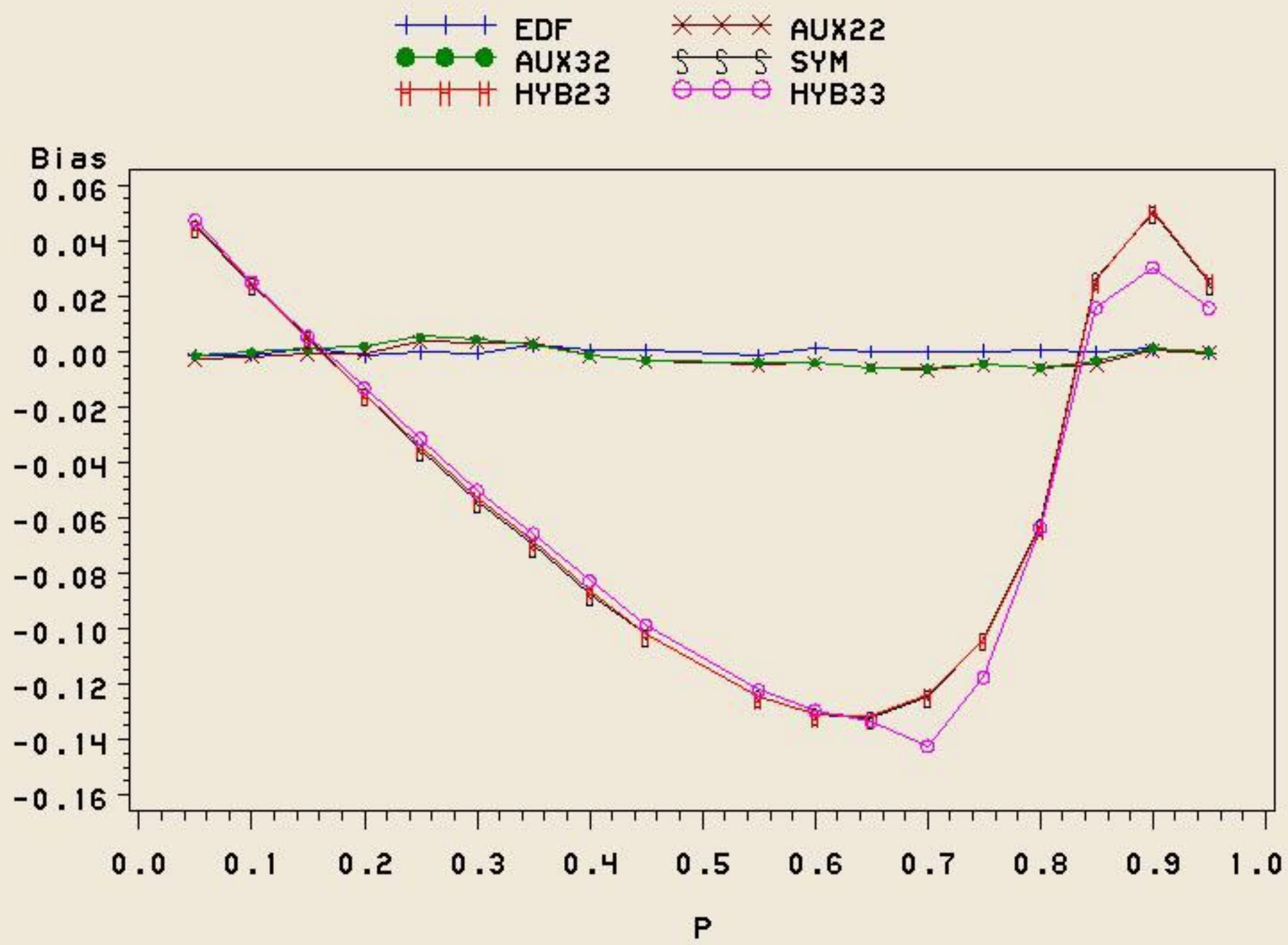


Figure 10: Relative Efficiency for $(X, Y) = (\text{Exp.}, \text{Normal})$ with $\theta = 100$ and $m = 2000$

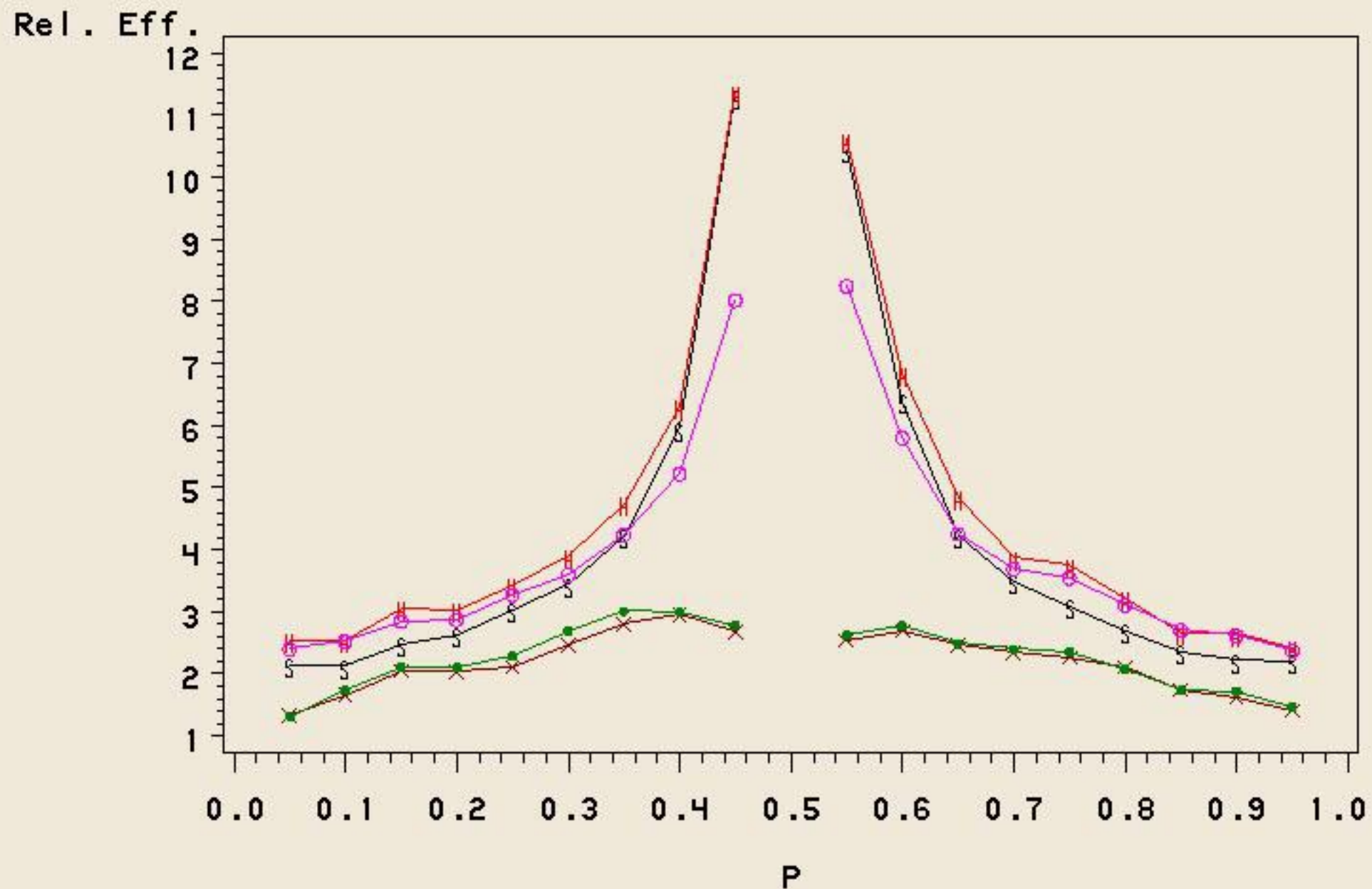
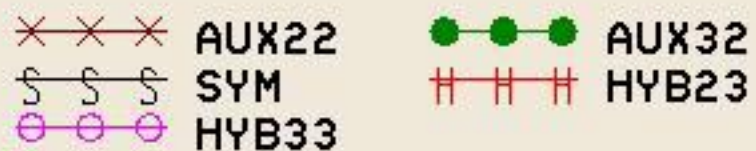


Figure 11: Relative efficiency for $(X,Y)=(Exp., Uniform)$ with $\theta=100$ and $m=2000$

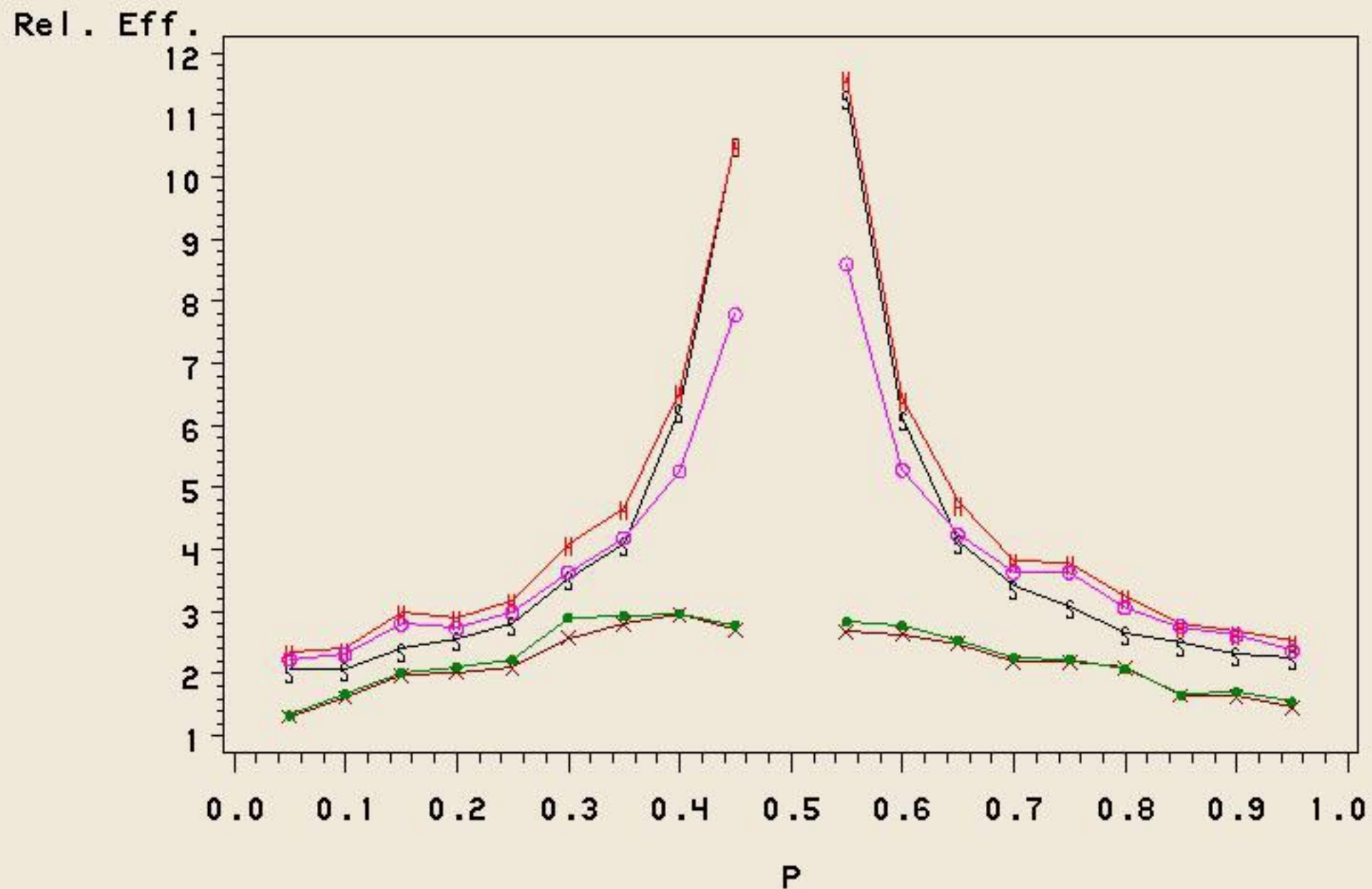
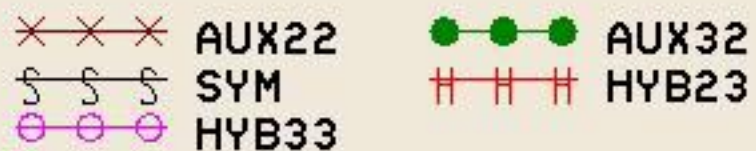


Figure 12: Relative Efficiency for $(X, Y) = (\text{Exp.}, \text{Exp.})$ with $\theta = 100$ and $m = 2000$

