Covariate Adjusted Correlation Analysis with Application to *FMR1* Premutation Female Carrier Data

**Damla Şentürk,[1] Danh V. Nguyen,[2*] Flora Tassone,[3]
Randi J. Hagerman,[4] Raymond J. Carroll[5] and Paul J. Hagerman[3]**

[1]Department of Statistics, Pennsylvania State University,
University Park, Pennsylvania 16802, U.S.A.
[2]Division of Biostatistics, University of California, Davis, California 95616, U.S.A.
[3]Department of Biochemistry and Molecular Medicine, University of California,
Davis, California 95616, U.S.A.
[4]Medical Investigation of Neurodevelopmental Disorders (M.I.N.D.) Institute,
University of California, Davis Health System, Sacramento, California 95817, U.S.A.
[5]Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
email: ucdnguyen@ucdavis.edu*

Summary

Motivated by molecular data on female premutation carriers of the fragile X mental retardation 1 (*FMR1*) gene, we present a new method of covariate adjusted correlation analysis to examine the association of messenger RNA (mRNA) and number of CGG repeat expansion in the *FMR1* gene. The association between the molecular variables in female carriers needs to adjust for activation ratio (ActRatio), a measure which accounts for the protective effects of one normal X chromosome in females carriers. However, there are inherent uncertainties in the exact effects of ActRatio on the molecular measures of interest. In order to account for these uncertainties, we develop a flexible adjustment that accommodates both additive and multiplicative effects of ActRatio nonparametrically. The proposed adjusted correlation uses local conditional correlations, which are local method of moments estimators, to estimate the Pearson correlation between two variables adjusted for a third observable covariate. The local method of moments estimators are averaged to arrive at the final covariate adjusted correlation estimator, which is shown to be consistent. We also develop a test to check the nonparametric joint additive and multiplicative adjustment form. Simulation studies illustrate the efficacy of the proposed method. Application to *FMR1* premutation data on 165 female carriers indicates that the association between mRNA and CGG repeat after adjusting for ActRatio is stronger. Finally, the results provide independent support for a specific jointly additive and multiplicative adjustment form for ActRatio previously proposed in the literature.

Key words: Conditional correlation; Fragile X syndrome; Local method of moments; Mental retardation; Nonparametric partial correlation; Pearson correlation; Semiparametric modeling.

1 Introduction

Fragile X syndrome (FXS) is the most common inherited form of X-linked intellectual disability, with cognitive and behavioral impairments associated with distinct physical features. FXS results from a hyperexpansion of a CGG trinucleotide repeat in the promoter region of the fragile X mental retardation 1 (*FMR1*) X-linked gene (Verkerk et al., 1991; Oberle et al., 1991). When the number of CGG repeats exceeds 200 (full mutation) methylation and transcriptional silencing of the gene occur (Pieretti et al., 1991) with consequent absence or deficiency of the FMR1 protein (FMRP; Devys et al., 1993). Individuals with smaller expansions in the premutation range of 55 to 200 CGG repeats are called premutation carriers. Many premutation carriers have some physical and behavioral characteristics of FXS (Hagerman, 2002) while a subgroup of older adult carriers develops fragile X-associated tremor/ataxia syndrome (FXTAS) later in their lives (Jacquemont et al., 2004) and about 20% develop premature ovarian failure. For a review, see Hagerman and Hagerman (2004). However, molecular mechanisms/models for the myriad of clinical involvements associated with premutation carriers, a current area of active research, are distinct from the molecular model that characterizes FXS. More precisely, unlike full mutation, premutation alleles do not lead to transcriptional silencing of *FMR1*. Indeed, it has been shown that premutation male carriers have significantly elevated levels of *FMR1* mRNA compared to normal controls (Tassone et al., 2000a, b; Kenneson et al., 2001) and mRNA levels are positively correlated with the number of CGG repeats.

For female premutation carriers, the underlying association/correlation between CGG repeat size and mRNA level is more complex. The analysis of this correlation needs to take into account (or adjust for) the protective effects from one normal X chromosome. This protective effect is quantified by the activation ratio (ActRatio), which measures the proportion of normal active X chromosomes. Although it is difficult to precisely account for the effect of ActRatio on observed mRNA level, Tassone et al. (2000a) proposed to examine the relationship between

CGG repeat size and mRNA level, after adjusting for ActRatio, based on the adjustment

$$\widetilde{X} = (1 - U)X + aU, \tag{1}$$

where $\widetilde{X}$ is the observed mRNA level, X is the unobserved (adjusted) mRNA level due to the carrier chromosome, U is the ActRatio, and a is the fixed mean level of mRNA in normal alleles. The parametric adjustment in (1) is a simple decomposition of the observed mRNA level into two parts, one from the normal allele and the other from the diseased allele. Although this simple decomposition serves as a simple and biologically sensible adjustment, it does not account for the inherent uncertainties in the precise effect of ActRatio on mRNA expression level (Tassone et al., 2000a). Hence, we propose a more general, fully nonparametric adjustment that incorporates both additive and multiplicative effects of U, as in (1). More precisely, we consider the following adjustment, of which the previous adjustment (1) is a special case,

$$\widetilde{X} = \phi_1(U)X + \phi_2(U), \tag{2}$$

where $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are unknown smooth functions of U. Similarly, the potential effect of U on the variability in CGG repeats is modeled as

$$\widetilde{Y} = \psi_1(U)Y + \psi_2(U), \tag{3}$$

where $\psi_1(\cdot)$ and $\psi_2(\cdot)$ are also allowed to be general unknown smooth functions to accommodate uncertainties in the effects of U, $\widetilde{Y}$ is the observed CGG repeat size and Y is its U-adjusted form. In (2)-(3), the unobserved variables (X, Y) are defined to be the parts of $(\widetilde{X}, \widetilde{Y})$ that are independent of U. Our aim is estimation of the correlation between X and Y, denoted ρ_{XY}, adjusted for the general effects of U based on the observed data $\{\widetilde{X}, \widetilde{Y}, U\}$.

The adjustments that we consider in (2)-(3) are flexible to accommodate linear effects of U, as in (1), or nonlinear effects. In addition, the effects of U may be additive, multiplicative or a combination of both. The trivial case where U has no effect is accommodated with $\phi_1(\cdot) =$

$\psi_1(\cdot) = 1$ and $\phi_2(\cdot) = \psi_2(\cdot) = 0$. Also, since there are no assumptions made on the unknown functions in (2)-(3), other than smoothness, an important property of the proposed estimator for ρ_{XY} is its invariance under linear transformations, similar to the Pearson correlation. Thus, the proposed covariate adjusted correlation is unaffected by the scale of the measurements.

We note that the adjustments (2)-(3) are partly related to the work of Şentürk and Müller (2005b). They proposed an estimator for ρ_{XY} (a) under the special case when $\widetilde{X} = X\phi(U)$ and $\widetilde{Y} = Y\psi(U)$ and (b) that requires identifiability conditions of $E\{\phi(U)\} = 1$ and $E\{\psi(U)\} = 1$ and the assumptions that $E(\widetilde{Y}) \neq 0$ and $E(\widetilde{X}) \neq 0$ for estimation. These identifiability conditions and assumptions are difficult to verify in practice generally, and they are not satisfied for the *FMR1* data adjustment described above. In the current work, we propose the more general adjustments (2)-(3) that account for the X-linked nature of disease and develop a completely different estimator for ρ_{XY} that does not require the above assumptions and identifiability conditions. A key observation in the development of the proposed estimator is that the (local) conditional correlation $Corr(\widetilde{X}, \widetilde{Y}|U = u)$ is equal to ρ_{XY} under (2)-(3), when $\phi_1(\cdot)$ and $\psi_1(\cdot)$ are of the same sign. This condition is satisfied when both $\phi_1(\cdot)$ and $\psi_1(\cdot)$ are positive. This means the observed measurements are positively correlated with what we want to measure, i.e. the correlation between X and $\widetilde{X}$ and between Y and $\widetilde{Y}$ are positive. Based on the above observation that $Corr(\widetilde{X}, \widetilde{Y}|U = u) = \rho_{XY}$ under (2)-(3), the data is stratified with respect to U, where the levels of U will be approximately constant in each stratum. We then average the local method of moments estimators of ρ_{XY}, obtained in each strata, to arrive at our final covariate adjusted correlation estimator.

Although the proposed adjustment formulation (2)-(3) is motivated from the problem of assessing the association between molecular measures, adjusted for ActRatio in female premutation carrier data, it is sufficiently general for a variety of other applications. Thus, the proposed adjusted estimator and the associated theory should be of broader interest beyond the motivating area of application. For instance, examples of covariate adjustments (2)-(3) include normalization of albumin turnover and protein catabolic rate (Kaysen et al., 2002), through division by U, body surface area. Such a normalization of the observed variables is common in biomedical studies, and can be viewed as a special case of the adjustments (2)-(3) with $\phi_2(\cdot) = \psi_2(\cdot) = 0$ and $\phi_1(\cdot) = \psi_1(\cdot) = U$. A similar adjustment in environmental health is described in Schisterman et al. (2005), where the exposure level of polychlorinated biphenyl (PCB), a lipophilic compound, is adjusted through division by a function of serum lipid levels (U).

We also note that although the additive effects of a covariate can be adjusted for with standard approaches, such as partial correlation or nonparametric partial correlation, these methods cannot adjust for multiplicative (possibly nonlinear) effects. A limitation of the partial correlation is that it adjusts for only additive linear effects of a covariate. More specifically, it can be shown that the standard partial correlation between $\widetilde{X}$ and $\widetilde{Y}$ adjusted for a covariate U targets ρ_{XY}, when $\widetilde{X} = X + a_1U + a_2$ and $\widetilde{Y} = Y + b_1U + b_2$. These standard methods no longer target ρ_{XY} under the more general adjustments (2)-(3). We elaborate on the special cases of the additive effects of U in Section 2 and the Appendix section.

The remainder of the article is organized as follows. We detail the proposed covariate adjusted estimator of ρ_{XY} in Section 2. The asymptotic result is also given in Section 2, where the proof is deferred to the Appendix section. In Section 3, we propose a bootstrap test to check the proposed dual additive and multiplicative adjustment structure of (2)-(3). The proposed method is further examined with simulation studies and illustrated with an application to the aforementioned data on female *FMR1* premutation carriers in sections 4 and 5, respectively.

2 Estimation

Estimation of ρ_{XY} is based on the observed data of size n, $\{(U_i, \widetilde{X}_i, \widetilde{Y}_i)\}_{i=1}^n$, where $\widetilde{X}_i = X_i\phi_1(U_i) + \phi_2(U_i)$, $\widetilde{Y}_i = Y_i\psi_1(U_i) + \psi_2(U_i)$ and the unobserved variables (X, Y) are defined to be the parts of $\widetilde{X}$ and $\widetilde{Y}$ that are independent of U. The proposed estimator of ρ_{XY} is constructed from local method of moments estimates of ρ_{XY}. These local estimates utilize the fact that, under the general adjustments (2)-(3), the correlation between $\widetilde{X}$ and $\widetilde{Y}$ at a fixed U is equal to the correlation ρ_{XY}. To be more precise, denote $\widetilde{\rho}(u)$ to be the correlation between $\widetilde{X}$ and $\widetilde{Y}$ given $U = u$, defined by $\widetilde{\rho}(u) \equiv Corr(\widetilde{X}, \widetilde{Y}|U = u) = Cov(\widetilde{X}, \widetilde{Y}|U = u)/\{Var(\widetilde{X}|U = u)Var(\widetilde{Y}|U = u)\}^{1/2}$. Note that by conditioning on $U = u$, it follows from the definitions of $\widetilde{Y}$ and $\widetilde{X}$ and the invariance of ρ_{XY} to linear transformations that

$$\widetilde{\rho}(u) = \rho_{XY},$$

if $\phi_1(u)$ and $\psi_1(u)$ are assumed to be of the same sign. The above relationship implies that within a neighborhood of u, the correlation between the observed variables $\widetilde{X}$ and $\widetilde{Y}$, denoted $\rho_{\widetilde{X}\widetilde{Y}}$, will target ρ_{XY} of interest. The proposed estimator of ρ_{XY}, based on this relationship, is an average of localized method of moments estimates of $\widetilde{\rho}(u)$.

In order to obtain the targeted local estimates, we bin the observed data with respect to U. The range of U is divided into m equidistant intervals, referred to as bins and denoted by $B_1, \ldots, B_m$. Let L_j denote the number of subjects falling into bin j, $1 \leq j \leq m$. To track the observations that fall into a given bin, bin-specific observations are marked by a prime. For example, data for subject k in bin j is $(U'_{jk}, \widetilde{X}'_{jk}, \widetilde{Y}'_{jk})$, for $1 \leq k \leq L_j$. We define the following local method of moments estimator of the correlation between $\widetilde{X}$ and $\widetilde{Y}$ within bin j,

$$r_j = \frac{M_{\widetilde{X}\widetilde{Y},j} - M_{\widetilde{X},j}M_{\widetilde{Y},j}}{\sqrt{M_{\widetilde{X}^2,j} - M^2_{\widetilde{X},j}}\sqrt{M_{\widetilde{Y}^2,j} - M^2_{\widetilde{Y},j}}},$$

where $M_{\widetilde{X}\widetilde{Y},j} = L_j{}^{-1}\sum_{k=1}^{L_j} \widetilde{X}'_{jk}\widetilde{Y}'_{jk}$, $M_{\widetilde{X},j} = L_j{}^{-1}\sum_{k=1}^{L_j} \widetilde{X}'_{jk}$, $M_{\widetilde{Y},j} = L_j{}^{-1}\sum_{k=1}^{L_j} \widetilde{Y}'_{jk}$, $M_{\widetilde{X}^2,j} = L_j{}^{-1}\sum_{k=1}^{L_j} \widetilde{X}'^2_{jk}$ and $M_{\widetilde{Y}^2,j} = L_j{}^{-1}\sum_{k=1}^{L_j} \widetilde{Y}'^2_{jk}$. Guidelines for choosing the total number of bins

m will be given in the simulation studies of Section 4. Since r_j targets ρ_{XY} for all $j = 1, \ldots, m$, a natural estimator of ρ_{XY} can be based on the average of $\{r_j\}_{j=1}^{m}$. Therefore, the proposed covariate adjusted correlation estimator of ρ_{XY} is

$$r = \sum_{j=1}^{m} \frac{L_j}{n} r_j, \tag{4}$$

which is a weighted average of the bin specific estimators. Note that the weights are proportional to the numbers of points in each bin. The covariate adjusted estimator, r, is consistent for ρ_{XY}, as given by the following result. The proof is deferred to the Appendix section.

Theorem 1. *Under the technical conditions given in the Appendix,*

$$r \;=\; \rho_{XY} + O_p(c_n),$$

where $c_n = \{n/\log(n)\}^{-1/3}$.

We emphasize here that the consistency of the covariate adjusted correlation estimator, r, holds under the general additive and multiplicative adjustments (2)-(3). However, as pointed out in the Introduction section and proven in the Appendix section, the special case of additive linear effects of U (i.e. $\widetilde{X} = X + a_1U + a_2$ and $\widetilde{Y} = Y + b_1U + b_2$) can be handled with standard partial correlation analysis. The partial correlation estimate is obtained by first regressing (1) $\widetilde{X}$ on U and (2) $\widetilde{Y}$ on U to obtain two sets of residuals. The partial correlation estimate is then obtained as the Pearson correlation between the two sets of residuals. In contrast to the additive linear case, the partial correlation does not target ρ_{XY} under general additive effects of U on X and Y, such as nonlinear effects. More specifically, consider $\widetilde{X} = X + \phi(U)$ and $\widetilde{Y} = Y + \psi(U)$, where $\phi(\cdot)$ and $\psi(\cdot)$ are unknown smooth functions of U that may be nonlinear. Under these general additive effects, it is also shown in the Appendix section that a simple generalization of the partial correlation, called nonparametric partial correlation, targets ρ_{XY}. The only difference between partial and nonparametric partial correlation is that, for the later, the two sets of residuals are obtained from nonparametric regressions of $\widetilde{X}$ on U and $\widetilde{Y}$ on U. Both partial and nonparametric partial regression do not target ρ_{XY} under the more general form of (2)-(3), as shown in the Appendix section.

We note that while r is based on an equidistant binning procedure, alternate binning approaches can be integrated to the estimation procedure proposed above. One alternative approach that we also explored is based on nearest neighbor binning. As pointed out earlier, for the equidistant binning used, B_j, $j = 1, \ldots, m$, are fixed and equidistant; however, the number of data points, L_j, falling into each bin is random. In nearest neighbor binning, the bin lengths and boundaries are random, but each bin contains the same number of observations, denoted by L. This alternate binning utilizes the nearest neighbor idea by first ordering the observed distortion values U_i, $i = 1, \ldots, n$, and then forming the $m = n/L$ number of bins by grouping the sets of L nearest neighbor values among the ordered set starting with the first L to the last. Once the bins are formed, the rest of the procedure is the same as explained for the case of equidistant binning. We compare the performance of the two binning procedures in more detail in Section 4.4 with respect to various distributions for U.

Also, upon the suggestion of the editor, we explored a variation on the proposed estimator in (4) by replacing the r_j's in (4) with their Fisher's z transformed values (i.e. $.5\{\ln(1 + r_j) - \ln(1 - r_j)\}$). Comparison of r with this variation is given in Section 4.5.

For inference, we use the bootstrap percentile method to form confidence intervals based on the proposed covariate adjusted estimator in the analysis of the female $FMR1$ premutation data. The estimated nonparametric density of the standardized 1000 bootstrap estimates of ρ_{XY} is given in Figure 2 (bottom panel), along with the standard normal density curve. The fitted density appear close to the standard normal density, indicating that the percentile bootstrap approximation is reasonable. The coverage of the proposed bootstrap percentile confidence intervals are examined through simulations reported in Section 4.3.

An important practical issue with the application of the proposed estimator is the adequacy of the assumed adjustment forms (2)-(3). Although these assumed dual additive and multiplicative adjustment forms are fairly general compared to the additive linear restriction of other methods like partial correlation, it is still of interest to check the adequacy of these forms. We address this issue next by developing a bootstrap test to check this assumption.

3 Assessing the Adjustment Model Assumption

The dual additive and multiplicative adjustment form of (2)-(3) imply that the local correlation $\widetilde{\rho}(u) = \rho_{XY}$ is free of u. Hence, under the null hypothesis, $H_0 : \widetilde{Y} = \psi_1(U)Y + \psi_2(U)$ and $\widetilde{X} = \phi_1(U)X + \phi_2(U)$, the scatter plot of the local correlation estimates should be randomly scattered around the constant ρ_{XY}. Even though this scatter plot, augmented with a scatter plot smoother, can provide an initial graphical check of the above assumption, we also develop a more formal assessment through a hypothesis testing procedure. The proposed test will be based on the local correlation estimates $\{r_j\}_{j=1}^m$ coming from each bin. Let $\{U_j^M\}_{j=1}^m$ denote the corresponding midpoints of the bins. Then a smooth fit to the scatter plot of $\{(U_j^M, r_j)\}_{j=1}^m$ is expected to be a constant function under H_0. We consider a smooth test as they are expected to be more powerful (Hart, 1997, p. 140). Reasonable test statistics quantify departures of the smooth estimator from a horizontal line at the sample mean of $\{r_j\}$. Alternatively, one can quantify the departures (from the horizontal line at zero) of the smooth estimator fitted to the centered scatter plot, $\{(U_j^M, r_j^C)\}_{j=1}^m$, where $r_j^C = r_j - m^{-1}\sum_j^m r_j$. Similar to the statistics proposed by Hart (1997) and Şentürk and Müller (2005a), we adopt as a measure of departure,

$$R_n = \frac{1}{m}\sum_{\ell=1}^{m} |\widehat{\rho}(U_\ell^M; h_T)|,$$

where $\widehat{\rho}(U_\ell^M; h_T) = \sum_{j=1}^m r_j^C w_j(U_\ell^M, h_T)$ is the linear smooth fitted to the centered scatter plot using the bandwidth h_T and weights $w_j(U_\ell^M, h_T)$, evaluated at U_ℓ^M.

An automatic data-based choice of the bandwidth parameter h_T that is fast to implement and that leads to good results, adopted from Rice (1984), is

$$h_T = \arg\min_h \{T(h)\} = \arg\min_h \left\{ \frac{(1/m)RSS(h)}{1 - 2tr(\mathbf{W}_h)/m} \right\}, \tag{5}$$

where $\mathbf{W}_h$ is an $m \times m$ matrix with (ℓ, j)th element $w_j(U_\ell^M; h)$, $RSS(h) = \|\mathbf{r}' - \widehat{\boldsymbol{\rho}}\|^2$ for $\mathbf{r}' = (r_1, \ldots, r_m)^{\mathrm{T}}$, and $\widehat{\boldsymbol{\rho}} = \{\widehat{\rho}(U_1^M, h), \ldots, \widehat{\rho}(U_m^M, h)\}^{\mathrm{T}}$.

We approximate the sampling distribution of R_n by the wild bootstrap, since the local estimators r_j are heteroscedastic with their variance dependent on U. The bootstrap samples have the form $\{(U_1^M, r_1^C V_1), \ldots, (U_m^M, r_m^C V_m)\}$, where V_j is sampled from the two-point distribution attaching masses $(\sqrt{5}+1)/2\sqrt{5}$ and $(\sqrt{5}-1)/2\sqrt{5}$ to the points $-(\sqrt{5}-1)/2$ and $(\sqrt{5}+1)/2$ (Davison and Hinkley, 1997, p. 272). The variables $\{r_j^C V_j\}$ have mean zero and crudely approximate the variance and skewness of the underlying distribution, since $(V_j)_{j=1}^m$ are independent and identically distributed random variables with mean zero and with variance and third moment equal to one. Properties of this test are studied in Section 4.3.

4 Simulation Studies

In this section we summarize the simulation studies conducted to examine (1) the finite-sample performance of the proposed covariate adjusted correlation estimator and its relative performance in comparison to no adjustment, parametric adjustment of Tassone et al. (2000a), partial correlation and nonparametric partial correlation, (2) the sensitivity of the proposed estimator to the choice of the number of bins m, (3) the power of the proposed bootstrap test for checking the dual additive and multiplicative adjustment forms and the coverage of the proposed bootstrap percentile confidence interval, (4) the performance of the two binning procedures (equidistant and nearest neighbor) and their robustness to the distribution of U and (5) the performance of the alternative estimator proposed via Fisher's z transformations.

4.1 Finite Sample Performance and Comparison to Other Adjustments

The simulation set-up was designed to reflect the observed *FMR1* premutation data, where the means and variation of ($\widetilde{X}$, $\widetilde{Y}$, U) are chosen to be similar to those of ($\widetilde{\text{mRNA}}$, $\widetilde{\text{CGG}}$, ActRatio). Also, the correlation $\rho_{\widetilde{X},\widetilde{Y}} = 0.32$ was chosen to approximately match the observed correlation $r_{\widetilde{\text{mRNA}},\widetilde{\text{CGG}}} = 0.29$. The covariate U is simulated from Uniform$[0.2, 0.9]$. The underlying unobserved variables $(X, Y)^{\text{T}}$ are obtained from the bivariate normal distribution with mean vector $(5.5, 7.5)^{\text{T}}$, $\sigma_X^2 = 1.5$, $\sigma_Y^2 = 1.1$ and $\rho_{XY} = 0.6$. The functional effects of U are given by $\psi_1(U) = (U + 4)^2/2$, $\psi_2(U) = 25(1 - U)$, $\phi_1(U) = U^3$ and $\phi_2(U) = 2(1 - U^2)$. The observed data is obtained as $\widetilde{X} = X\phi_1(U) + \phi_2(U)$ and $\widetilde{Y} = Y\psi_1(U) + \psi_2(U)$. The simulation studies were carried out for sample sizes of $n = 150, 300$ and 600. The proposed covariate adjusted correlation estimator is compared to estimators from no adjustment ($r_{\widetilde{X}\widetilde{Y}}$), the parametric adjustment (1) of Tassone et al. (2000a), partial correlation and nonparametric partial correlation under three distortion settings: (a) nonparametric additive and multiplicative distortion effects, (b) parametric additive and multiplicative distortion as given in equation (1) and (c) parametric additive distortion.

The first distortion considered is (a) the general case of nonparametric additive and multiplicative effects under which only the proposed covariate adjusted estimator targets the underlying correlation coefficient. Table 2 reports the estimated absolute bias, variance and MSE of the correlation estimators based on 1000 Monte Carlo data sets for each sample size. As evident from the results in Table 2, only the bias of the proposed covariate adjusted correlation decreases with increasing sample size. As expected, the biases of the other methods remain substantial across the different sample sizes, since they do not target ρ_{XY}.

In the second distortion set-up (b) of parametric additive and multiplicative distortion ($\widetilde{Y} = Y$, $\widetilde{X} = (1 - U)X + 1.42U$), the parametric adjustment (1) of Tassone et al. and the proposed covariate-adjusted estimator are the two methods that target the underlying correlation. For the third set-up (c) of parametric additive distortion ($\widetilde{Y} = Y + 5U$, $\widetilde{X} = X + 10U$), partial correlation, nonparametric partial correlation and the proposed method target the correct correlation. The results for (b) and (c) are reported in Tables 3 and 4, respectively. As can be seen from Table 3, both partial and nonparametric partial correlation perform (equally) poorly and their biases do not decrease with increasing sample size, as expected for model (b). For parametric additive distortion, namely case (c), the incorrect form of adjustment (1) due to Tassone et al. and the unadjusted correlation result in biases that do not decrease with increasing sample size (Table 4). The biases of parametric, nonparametric partial correlation and the proposed covariate-adjusted correlation decrease with increasing n. The simpler methods of parametric and partial correlation are more efficient than the proposed method under the null models (b), (c) for the small sample size of $n = 150$, as expected. However, this difference seems to diminish quickly as the sample size increases to $n = 300$ and 600.

4.2 Choice of m

In the simulation studies we also examine the effect of the total number of bins, m, on the proposed estimators. Similar to the results of Şentürk and Müller (2005b), where the corresponding estimator was also obtained through binning, the estimates are found to be robust to the choice of m. The results indicate that the correlation estimates and MSEs were very similar for m between 15-30 for $n = 150$, m between 20-40 for $n = 300$ and m between 25-50 for $n = 600$. For example, under simulation set-up (a) of Section 4.1, for $n = 300$ with $m \in \{20, 30, 40\}$ the mean of covariate adjusted correlation estimates for $\rho_{XY} = 0.467$ were (0.449, 0.447, 0.439) and the MSEs were (0.0027, 0.0030, 0.0034), respectively. Although our experience with the proposed estimator indicates that the final estimate is fairly robust to a reasonably wide range of m in practice, for any given application it is prudent to consider an analysis of sensitivity to m, as was done for the *FMR1* data application above.

4.3 Power of the Proposed Bootstrap Test and Coverage of the Proposed Bootstrap Confidence Interval

Next, to examine the power of the proposed bootstrap test for checking the dual additive and multiplicative form of (2)-(3), we considered two cases of deviations (alternatives) from this assumption (null case). The simulation model (a) of Section 4.1 described above are used for the null case. In the first alternative case, $\widetilde{Y}$ and $\widetilde{X}$ deviate from the additive and multiplicative forms (2)-(3) through:

$$\begin{aligned} \widetilde{X} &= N_0 - N_0 I_{\{\theta>0\}} + \cos\{X(0.4+\theta/130)(U/1.3+3.2)\} I_{\{\theta>0\}} \\ \widetilde{Y} &= M_0 - M_0 I_{\{\theta>0\}} + \cos\{Y(0.4+\theta/130)(U/1.3+3.2)\} I_{\{\theta>0\}}, \end{aligned}$$

where $N_0 \equiv \phi_1(U)X + \phi_2(U)$, $M_0 \equiv \psi_1(U)Y + \psi_2(U)$, $I_{\{E\}}$ is the indicator function for event E, and $\theta = 0, 1, \ldots, 8$. The functions $\phi_1(U)$, $\phi_2(U)$, $\psi_1(U)$ and $\psi_2(U)$ are as defined above. The null hypothesis (i.e. assumption (2)-(3) holds) corresponds to $\theta = 0$ and $\theta = 1, \ldots, 8$ correspond to increasing alternatives. These alternatives as well as the null are displayed in Figure 3 (top left plot) where the conditional correlation functions, $\tilde{\rho}(u)$, are provided. When the additive and multiplicative forms are satisfied ($\theta = 0$), $\tilde{\rho}(u)$ is constant (see Section 3). The second set of alternatives/violations explored in this simulation study are provided graphically in the top right plot of Figure 3. These conditional correlation functions correspond to the following alternative deviations:

$$\begin{aligned} \widetilde{X} &= N_0 - N_0 I_{\{\theta>0\}} + \cos\{X(0.38+\theta/130)(2U-1.1)\} I_{\{\theta>0\}}, \\ \widetilde{Y} &= M_0 - M_0 I_{\{\theta>0\}} + \cos\{Y(0.38+\theta/130)(2U-1.1)\} I_{\{\theta>0\}}, \end{aligned}$$

for $\theta = 0, \ldots, 8$. Similarly, the null hypothesis corresponds to the case of $\theta = 0$.

Given in the lower panel of Figure 3 are the power of the bootstrap test proposed in Section 3 to check the adequacy of the dual additive and multiplicative forms. Displayed are three sets of power curves corresponding to sample sizes $n = 150, 300$ and 600. Two curves of the same line type are for the test at significance levels 0.05 (bottom curve) and 0.10 (top curve). Power estimates are based on 1000 Monte Carlo runs. The observed type I errors at $\theta = 0$, for the above significance levels are, (0.015, 0.025) for $n = 150$, and (0.04, 0.08) for $n = 600$ in the first alternative deviation case. For the second alternative deviation case, the observed levels are (0.01, 0.04) for $n = 150$ and (0.04, 0.09) for $n = 600$. As expected, the levels of the bootstrap test move closer to the target values and the power functions increase with increasing deviation away from the null case of $\theta = 0$ and with increasing sample size.

We also examined the estimated coverage levels of the proposed bootstrap percentile confidence intervals under the simulation setting (a) of Section 4.1. Briefly, one thousand data sets were simulated at two sample sizes of $n = 165$ and $n = 300$. For each data set, 1000 bootstrap samples were generated and the estimated coverage values of the CIs corresponding to levels of (0.80, 0.90 and 0.95) are (0.75, 0.88, 0.93) for $n = 165$ and (0.80, 0.89, 0.95) for $n = 300$.

4.4 Alternate Binning Procedure and Robustness to the Distribution of U

We also ran a simulation study to compare the proposed equidistant binning estimator to one obtained via nearest neighbor binning and to evaluate the performance of the two binning procedures under different U distributions. While the bin size is kept constant in equidistant binning, the number of points per bin is kept constant in nearest neighbor binning. (See Section 2.) We compare the two binning procedures for three different distributions of U. Under model (a) of Section 4.1 where U was sampled from Uniform$[0.2, 0.9]$ uniform distribution, we additionally consider cases where U is sampled from $\mathcal{N}(0.55, 0.04)$ and $\chi^2(1)/6+0.4$. The three distributions are chosen such that they have approximately the same first two moments. The results are summarized in Table 5. The results suggest that the proposed equidistant binning is quite robust to the distribution of U and the nearest neighbor binning do not improve on the proposed equidistant binning.

4.5 Comparison to Estimators via Fisher's z Transformation

We implemented the variation of the proposed estimator by replacing r_j by its the Fisher-z values, and compared its performance to that of r under the simulation set-up of Section 4.4 in terms of bias, variance and MSE. We compared their performance under U distributed as uniform, normal and χ^2 as described earlier. Even though neither estimation approach was superior to the other in all aspects, we believe the simulation results reported in Table 6 are of interest. While the performance of the estimators are quite similar for uniformly distributed U, the estimator averaging Fisher-z values improves on the bias of the proposed estimator for U distributed as normal and χ^2. However, the original proposed estimator yields smaller variance, especially for small sample sizes ($n = 150$ and 300); thus original estimator results in smaller MSE for the smaller sample size. Their MSE estimates are similar for larger n (300 and 600) in the simulation study, as expected.

5 Application to Female *FMR1* Premutation Data

The molecular measurements, $\widetilde{\mathrm{CGG}}, \widetilde{\mathrm{mRNA}}$ and activation ratio $U \equiv \mathrm{ActRatio}$, were obtained from experiments at the University of California at Davis on 165 female premutation carriers. Our main interest here is to target $\rho_{XY} \equiv \rho_{\mathrm{mRNA,CGG}}$, the activation ratio-adjusted correlation between the mRNA level and CGG repeat size. Figure 1 gives the matrix plot for the observed variables $[\widetilde{\mathrm{CGG}}, \widetilde{\mathrm{mRNA}}, \mathrm{ActRatio}]$, where $\widetilde{\mathrm{CGG}}$, $\widetilde{\mathrm{mRNA}}$ and ActRatio range between $(57, 138)$ repeats, $(0.78, 6.3)$ and $(0.19, 0.91)$, respectively. Figure 1 suggests a potential ActRatio effect on both $\widetilde{\mathrm{CGG}}$ and $\widetilde{\mathrm{mRNA}}$.

Prior to estimating $\rho_{\mathrm{mRNA,CGG}}$, we assess the adequacy of the assumed dual additive and multiplicative forms (2)-(3) for the data, as described in Section 3. The local correlation estimators from each bin are given in Figure 2 (top panel), along with a local linear smooth using the automatic bandwidth choice of $h = 0.1$ determined by (5). A p-value of 0.56 was obtained from 1000 bootstrap replications of R_n. Thus, the adequacy of the assumed adjustment forms (2)-(3) is not rejected. Graphically, this can also be seen from Figure 2 where the linear smooth fitted to the scatter plot of the local correlations is approximately close to a constant function.

In our analysis, we compare the proposed covariate (ActRatio) adjusted estimate for the correlation, $\rho_{\text{mRNA,CGG}}$, to estimates obtained without adjustment and with adjustment (1) on $\widetilde{\text{mRNA}}$, previously proposed by Tassone et al. (2002a), partial correlation and nonparametric partial correlation. The estimate without adjustment corresponds to the observed Pearson correlation between $\widetilde{\text{mRNA}}$ and $\widetilde{\text{CGG}}$ ($\widetilde{X}$ and $\widetilde{Y}$). The proposed estimate is obtained using a total of 20 bins. We note here that the covariate adjusted correlation estimate was quite robust to the choice of the number of bins. For example, the estimates were very similar for the number of bins from $m = 17$ to $m = 25$. The estimates and approximate 95% confidence intervals (CIs) for ρ_{XY} from these five methods are provided in Table 1. For the unadjusted Pearson correlation, the assumed adjustment (1), partial correlation and nonparametric partial correlation approximate confidence intervals (CIs) can be obtained using Fisher's z-transformation. The CI for the proposed covariate adjusted estimator in (4) was obtained using the percentile bootstrap method with 1000 bootstrap replications.

The correlation between the observed mRNA levels and the CGG in female premutation carriers, unadjusted for the effect of activation ratio, is 0.29 (95% CI: $0.15 - 0.43$, Table 1). Although still significant, the correlation estimate for female carriers falls substantially below the corresponding estimate for male carriers (correlation ≈ 0.57), which has been established in literature. As described in the Introduction section, this weaker association in female carriers is attributed partly to the protective effects from one normal X chromosome in female carriers which is absent in male carriers. Applying the (parametric) adjustment (1) of Tassone et al. (2000a), specifically $\widetilde{\text{mRNA}} = (1-\text{ActRatio})\text{mRNA}+a\text{ActRatio}$, to account for activation ratio results in a stronger (adjusted) correlation point estimate of 0.34. We used the constant $a =$

1.42, which is the empirical mean mRNA level for normal/unaffected individuals from Tassone et al. (2000b). The proposed covariate adjusted correlation under the general adjustment forms (2)-(3), allowing for nonparametric effects of ActRatio on both observed mRNA and CGG, results in an adjusted correlation point estimate of 0.37 (95% CI: $0.25 - 0.51$). Although the proposed method suggests that the underlying correlation is slightly higher, this result is quite similar to the result using the parametric additive and multiplicative adjustment (1) proposed by Tassone et al. (2000a). Hence, this application provides an independent empirical support for the previously proposed parametric joint additive and multiplicative effect of ActRatio on mRNA, derived mainly from biological motivations. Also, since the nonparametric partial correlation (0.367, 95% CI: $(0.228, 0.491)$) is close to the proposed adjusted correlation estimate (0.372, 95% CI: $(0.252, 0.520)$), informally, it is interpreted that the nonlinearity is due to the additive distortion part.

6 Concluding Remarks

Motivated by an adjustment for activation ratio in fragile X premutation female carriers, we proposed a general dual additive and multiplicative correlation adjustment model for the correlation between mRNA and CGG repeat expansion. A key feature of the methodology is that the uncertainty in the precise effects of activation ratio at the molecular level is modeled nonparametrically, thus, accommodating linear or nonlinear effects. We proposed a simple covariate adjusted correlation estimator that is easy to obtain, showed that it is consistent, and examined its numerical properties in simulation studies. Although the adjustment forms are fairly general and, therefore, are automatically adaptive to special cases like linear additive or nonlinear additive effects, we also developed and assessed the performance of a bootstrap test procedure to check the adequacy of the dual additive and multiplicative forms. A test for detecting whether the distortion setting at hand would reduce to parametric or only additive cases would also be of interest, since then a simpler adjustment method can be employed. Nevertheless this remains an open problem requiring further research.

Application of the proposed covariate adjusted correlation to $n = 165$ fragile X premutation female carriers indicates stronger association between *FMR1* mRNA level and CGG repeat expansion compared to unadjusted analysis. Our results provide new insights and additional support for a dual additive and multiplicative parametric adjustment previously proposed in the fragile X premutation literature. The proposed adjustment is also applicable to the multiplicative adjustments (normalizations) used in biomedical research, including adjustments of biomarkers of inflammation by body mass index or body surface area and individual levels of PCB exposure by individual serum lipids.

Extension of the proposed algorithm to accommodate multiple covariates poses challenges. While the adjustment for two covariates ($\mathbf{U} = (U_1, U_2)$) would be a straight forward extension of the proposed algorithm using a two dimensional binning procedure, as the dimension of $\mathbf{U}$ increases, one would quickly run into the curse of dimensionality. Since the proposed procedure involves localizing with respect to $\mathbf{U}$, when the dimension of $\mathbf{U}$ increases, the data needed for the localization (binning) would become highly sparse. In these cases, a dimension reduction approach, such as taking a linear combination of the components of $\mathbf{U}$ vector may be of interest.

Acknowledgment

We are grateful to the reviewer, Associate Editor and the Editor for many detailed suggestions which substantially improved the paper. Support for this work includes the National Institute of Health grants UL1RR024922, RL1AG032119 and RL1AG032115 (DVN, FT, RJH, PJH), National Institute of Child Health and Human Development grant HD036071(RJH, DVN, FT), National Cancer Institute grant CA-57030 (RJC) and grant UL1 RR024146 from the National Center for Research Resources (NCRR).

References

Davidson, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Applications.* New York: Cambridge University Press.

Devys, D., Lutz, Y., Rouyer, N., Bellocq, J. P., Mandel, J. L. (1993). The FMR-1 protein is cytoplasmic, most abundant in neurons and appears normal in carriers of a fragile X premutation. *Nature Genetics* **4**, 335–340.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* London: Chapman and Hall.

Hagerman, R. J. (2002). Physical and behavioral phenotype. In *Fragile X syndrome: Diagnosis, treatment and research*, 3rd ed, R. J. Hagerman and P. J. Hagerman, 3–109. Baltimore: Johns Hopkins University Press.

Hagerman, P. J. and Hagerman, R. J. (2004). The fragile-X premutation: a maturing perspective. *American Journal of Human Genetics* **74**, 805–816.

Härdle, W., Janssen, P. and Serfling, R. (1988). Strong uniform consistency rates for estimators of conditional functionals. *Annals of Statistics* **16**, 1428–1449.

Hart, J. (1997). *Nonparametric Smoothing and Lack of Fit Tests.* New York: Springer-Verlag.

Jacquemont, S., Hagerman, R. J., Leehey, M. A., Hall, D. A., Levine, R. A., Brunberg, J. A., Zhang, L., Jardini, T., Gane, L. W., Harris, S. W., Herman, K., Grigsby, J., Greco, C., Berry-Kravis, E., Tassone, F. and Hagerman, P. J. (2004). Penetrance of the fragile X-associated tremor/ataxia syndrome (FXTAS) in a premutation carrier population: Initial results from the California-based study. *Journal of the American Medical Association* **291**, 460–469.

Kaysen, G. A., Dubin, J. A., Müller, H. G., Mitch, W. E., Rosales, L. M., Levin, N. W. and the Hemo Study Group (2002). Relationship among inflammation nutrition and physiologic mechanisms establishing albumin levels in hemodialysis patients. *Kidney International* **61**, 2240–2249.

Kenneson, A., Zhang, F., Hagedorn, C. H. and Warren, S. T. (2001). Reduced FMRP and increased FMR1 transcription is proportionally associated with CGG repeat number in intermediate-length and premutation carriers. *Human Molecular Genetics* **10**, 1449–1454.

Oberlé, L., Rousseau, F., Heitz, D., Kretz, C., Devys, D., Hanauer, A., Boue, J., Bertheas, M. F. and Mandel, J. L. (1991). Instability of a 550-base pair DNA segment and abnormal methylation in fragile X sydrome. *Science* **252**, 1097–1102.

Pieretti, M., Zhang, F., Fu, Y.-H., Warren, S. T., Oostra, B. A., Caskey, C. T. and Nelson, D. L. (1991). Absence of expression of the FMR-1 gene in fragile X syndrome. *Cell* **66**, 817–822.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics* **12**, 1215–1230.

Schisterman, E. F., Whitcomb, B. w., Louis, G. M. B. and Louis, T. A. (2005). Lipid adjustment in the analysis of environmental contaminants and human health risks. *Environmental Health Perspectives* **113**, 853–857.

Şentürk, D. and Müller, H. G. (2005a). Covariate adjusted regression. *Biometrika* **92**, 59–74.

Şentürk, D. and Müller, H. G. (2005b). Covariate adjusted correlation analysis via varying coefficient models. *Scandinavian Journal of Statistics* **32**, 365–383.

Tassone, F., Hagerman, R. J., Chamberlain, W. D. and Hagerman, P. J. (2000a). Transcription of the FMR1 gene in individuals with fragile X syndrome. *American Journal of Medical Genetics* **97**, 195–203.

Tassone, F., Hagerman, R. J., Taylor, A. K., Gane, L. W., Godfrey, T. E. and Hagerman, P. J. (2000b). Elevated levels of FMR1 mRNA in carrier males: a new mechanism of involvement in fragile X syndrome. *American Journal of Human Genetics* **66**, 6–15.

Verkerk, A. J., Pieretti, M., Sutcliffe, J. S., Fu, Y. H., Kuhl, D. P., Pizzuti, A., Reiner, O., Richards. S., Victoria, M. F., Zhang, F., Eussen, B. E., van Ommen, G.-J.B., Blonden, L. A. J., Riggins, G. J., Chastain, J. L., Kunst, C. B., Galjaard, H., Caskey, C. T., Nelson, D. L., Oostra, B. A. and Warren, S. T. (1991). Identification of a gene (FMR-1) containing a CGG repeat co-incident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914.

Appendix

Proof of Consistency

We first state the technical conditions that will be used in the proof of consistency. They are: *C1.* The adjusting variable U is independent of the variables X and Y. In addition, the marginal density $f(U)$ of U has compact support, i.e. $a \leq U \leq b$ for some constants a , b, and satisfies $\inf_{a\leq u\leq b} f(u) > 0$, $\sup_{a\leq u\leq b} f(u) < \infty$. The marginal density is also uniformly Lipschitz. *C2.* The functions $\phi_1(\cdot)$, $\phi_2(\cdot)$, $\psi_1(\cdot)$ and $\psi_2(\cdot)$ have continuous derivatives. Furthermore, $\phi_1(\cdot)$ and $\psi_1(\cdot)$ are of the same sign. *C3.* The two variables X and Y satisfy the following moment conditions, $E|X|^{2\lambda} < \infty$ and $E|Y|^{2\lambda} < \infty$ for some $\lambda \geq 3$. *C4.* The function $h(u) = \int xg(x,u)dx$ is uniformly Lipschitz, where $g(\cdot,\cdot)$ is the joint density function of $\widetilde{X}$ and U and of $\widetilde{Y}$ and U. *C5.* The number of bins m is of order $\{n/\log(n)\}^{1/3}$.

Considering the definition of the Nadaraya-Watson estimator (Fan and Gijbels, 1996), we note that all the five terms in r_j are Nadaraya-Watson estimators. For instance, consider $M_{\widetilde{X},j}$. It has the form $M_{\widetilde{X},j} = L_j^{-1} \sum_{k=1}^{L_j} \widetilde{X}'_{jk} = \{\sum_{i=1}^{n} K((U_i - U_j^M)/h)\widetilde{X}_i\}/\sum_{i=1}^{n} K((U_i - U_j^M)/h) \equiv \hat{N}(U_j^M)$, which is a Nadaraya-Watson estimator with $K(\cdot) = (1/2)\mathbf{1}_{[-1,1]}$, $h = (b-a)/m$, and U_j^M is the midpoint of the jth bin, as defined in Section 3. Uniform consistency of Nadaraya-Watson estimators with kernels of compact support has been shown in Härdle *et al.* (1988),

$$\sup_{a \le u \le b} |\hat{N}(u) - N(u)| = O_p(c_n), \tag{6}$$

where $N(u) = E(\widetilde{X}|U = u)$ and $c_n = \{n/\log(n)\}^{-1/3}$. Then (6) implies $\sup_j |\hat{N}(U_j^M) - N(U_j^M)| = O_p(c_n)$. Similar to the uniform consistency of $M_{\widetilde{X},j}$, it follows that $M_{\widetilde{Y},j}$, $M_{\widetilde{X}^2,j}$, $M_{\widetilde{Y}^2,j}$ and $M_{\widetilde{X}\widetilde{Y},j}$ are all uniformly consistent over j. Hence the following holds uniformly in j

$$\begin{aligned} r_j &= \frac{E(\widetilde{X}\widetilde{Y}|U = U_j^M) - E(\widetilde{X}|U = U_j^M)E(\widetilde{Y}|U = U_j^M)}{\sqrt{E(\widetilde{X}^2|U = U_j^M) - \{E(\widetilde{X}|U = U_j^M)\}^2}\sqrt{E(\widetilde{Y}^2|U = U_j^M) - \{E(\widetilde{Y}|U = U_j^M)\}^2}} + o_p(c_n) \\ &= Corr(\widetilde{X}, \widetilde{Y}|U = U_j^M) + o_p(c_n) = \rho_{XY} + o_p(c_n). \end{aligned}$$

Hence, Theorem 1 follows.

Remark: The consistency of r also follows under weaker moment conditions than the ones given in *C3*, where $2 < \lambda < 3$. For the order of m and the convergence rates under weaker moment conditions, see Härdle *et al.* (1988) for details.

Target of partial correlation

The partial correlation between $\widetilde{Y}$ and $\widetilde{X}$ adjusted for U is equivalent to the correlation between the variables $e_{\widetilde{Y}U}$ and $e_{\widetilde{X}U}$, denoted $\rho_{e_{\widetilde{Y}U}e_{\widetilde{X}U}}$, where $e_{\widetilde{Y}U}$ and $e_{\widetilde{X}U}$ are the errors from the regression models $\widetilde{Y} = a_0 + a_1 U + e_{\widetilde{Y}U}$ and $\widetilde{X} = b_0 + b_1 U + e_{\widetilde{X}U}$, respectively. Using the population normal equations for regression, under adjustments (2)-(3), we have $a_1 = \text{cov}\{U, \widetilde{Y}\}/\text{var}(U) = [\text{cov}\{U, \psi_2(U)\} + \text{cov}\{U, \psi_1(U)Y\}]/\text{var}(U)$, $a_0 = E(\widetilde{Y}) - a_1 E(U)$, $b_1 = \text{cov}\{U, \widetilde{X}\}/\text{var}(U) = [\text{cov}\{U, \phi_2(U)\} + \text{cov}\{U, \phi_1(U)X\}]/\text{var}(U)$, and $b_0 = E(\widetilde{X}) - b_1 E(U)$.

Thus, plugging in definitions of a_0 and b_0, we have $e_{\widetilde{Y}U} = \widetilde{Y} - a_0 - a_1 U = [\psi_2(U) - E\{\psi_2(U)\}] + [\psi_1(U)Y - E\{\psi_1(U)Y\}] - a_1\{U - E(U)\}$ and $e_{\widetilde{X}U} = \widetilde{X} - b_0 - b_1 U = [\phi_2(U) - E\{\phi_2(U)\}] + [\phi_1(U)X - E\{\phi_1(U)X\}] - b_1\{U - E(U)\}$. Therefore, $\rho_{e_{\widetilde{Y}U} e_{\widetilde{X}U}}$ is not necessarily equal ρ_{XY}. However note that $\rho_{e_{\widetilde{Y}U} e_{\widetilde{X}U}} = \rho_{XY}$ holds if $\psi_1(U) = \phi_1(U) = 1$ and $\psi_2(U)$ and $\phi_2(U)$ are both linear in U, i.e. $\widetilde{X} = X + c_1 U + c_2$ and $\widetilde{Y} = Y + d_1 U + d_2$. This follows from the fact that $a_1 = c_1$ and $b_1 = d_1$ and hence $e_{\widetilde{Y}U} = Y - EY$ and $e_{\widetilde{X}U} = X - EX$ when $\psi_1(U) = \phi_1(U) = 1$ and $\psi_2(U)$ and $\phi_2(U)$ are linear in U. An implicit assumption here is that $\mathrm{cov}(X, U) = \mathrm{cov}(Y, U) = 0$.

Target of nonparametric partial correlation

The nonparametric partial correlation between $\widetilde{Y}$ and $\widetilde{X}$ adjusted for U is equivalent to $\rho_{\tilde{e}_{\widetilde{Y}U} \tilde{e}_{\widetilde{X}U}}$, where $\tilde{e}_{\widetilde{Y}U}$ and $\tilde{e}_{\widetilde{X}U}$ are the errors from the nonparametric regression models $\widetilde{Y} = E(\widetilde{Y}|U) + \tilde{e}_{\widetilde{Y}U}$ and $\widetilde{X} = E(\widetilde{X}|U) + \tilde{e}_{\widetilde{X}U}$, respectively. Thus, under adjustments (2)-(3), $\tilde{e}_{\widetilde{Y}U} = \widetilde{Y} - E(\widetilde{Y}|U) = \psi_1(U)\{Y - E(Y|U)\}$ and $\tilde{e}_{\widetilde{X}U} = \widetilde{X} - E(\widetilde{X}|U) = \phi_1(U)\{X - E(X|U)\}$. Therefore, $\rho_{\tilde{e}_{\widetilde{Y}U} \tilde{e}_{\widetilde{X}U}}$ does not necessarily equal ρ_{XY} unless $\psi_1(U) = \phi_1(U) = 1$, i.e. $\widetilde{X} = X + \phi(U)$ and $\widetilde{Y} = Y + \psi(U)$ and X and Y are independent of U.

Table 1: Estimates and approximate 95% confidence intervals (CIs) for $\rho_{\text{mRNA,CGG}}$ adjusted for ActRatio in $n = 165$ female premutation carriers. The first four estimates correspond to unadjusted Pearson correlation and parametric adjustment (1) from Tassone et al. (2000a), partial correlation and nonparametric partial correlation with approximate CIs obtained using Fisher's z-transformation. CIs for the proposed covariate adjusted correlation, adjusting for ActRatio, were obtained from the bootstrap percentile method.

Estimation Method	**Lower Limit**	**Point Estimate**	**Upper Limit**
Unadjusted Pearson correlation $r_{\widetilde{X}\widetilde{Y}}$	0.150	0.295	0.427
Parametric adjustment (1)	0.201	0.343	0.471
Partial correlation	0.218	0.357	0.483
Nonparametric partial correlation	0.228	0.367	0.491
Covariate adjusted correlation	0.252	0.372	0.520

Table 2: Estimated absolute bias, variance and MSE of the estimators from proposed covariate adjusted correlation, no adjustment, parametric adjustment (1) of Tassone et al. (2000a), partial correlation, and nonparametric partial correlations obtained under model (a) (nonparametric additive and multiplicative effects) of Section 4.1 based on 1000 Monte Carlo data sets. The results are presented for three sample sizes of $n = 150, 300$ and 600.

$n = 150$			
Estimation Method	**Bias**	**Variance**	**MSE**
Unadjusted Pearson correlation $r_{\widetilde{X}\widetilde{Y}}$	0.154	0.006	0.030
Parametric adjustment (1)	0.240	0.007	0.065
Partial correlation	0.158	0.007	0.032
Nonparametric partial correlation	0.099	0.006	0.016
Covariate adjusted correlation	0.040	0.006	0.008
$n = 300$			
Unadjusted Pearson correlation $r_{\widetilde{X}\widetilde{Y}}$	0.151	0.003	0.026
Parametric adjustment (1)	0.235	0.004	0.059
Partial correlation	0.150	0.003	0.026
Nonparametric partial correlation	0.092	0.003	0.011
Covariate adjusted correlation	0.020	0.003	0.003
$n = 600$			
Unadjusted Pearson correlation $r_{\widetilde{X}\widetilde{Y}}$	0.151	0.002	0.024
Parametric adjustment (1)	0.236	0.002	0.057
Partial correlation	0.151	0.002	0.024
Nonparametric partial correlation	0.092	0.001	0.010
Covariate adjusted correlation	0.013	0.001	0.001

Table 3: Estimated absolute bias, variance and MSE of the estimators of the correlation under the null case of parametric adjustment given in (1) from proposed covariate adjusted correlation, no adjustment, parametric adjustment (1) of Tassone et al. (2000a), partial correlation, and nonparametric partial correlations under model (b) (parametric additive and multiplicative effects) of Section 4.1 based on 1000 Monte Carlo data sets. The results are presented for three sample sizes of $n = 150, 300$ and 600.

$n = 150$ **Estimation Method**	**Bias**	**Variance**	**MSE**
Unadjusted Pearson correlation $r_{\tilde{X}\tilde{Y}}$	0.216	0.006	0.052
Parametric adjustment (1)	10^{-4}	0.004	0.004
Partial correlation	0.041	0.004	0.006
Nonparametric partial correlation	0.042	0.004	0.006
Covariate adjusted correlation	0.037	0.006	0.007
$n = 300$			
Unadjusted Pearson correlation $r_{\tilde{X}\tilde{Y}}$	0.215	0.003	0.049
Parametric adjustment (1)	0.002	0.002	0.002
Partial correlation	0.038	0.002	0.004
Nonparametric partial correlation	0.039	0.002	0.004
Covariate adjusted correlation	0.020	0.003	0.003
$n = 600$			
Unadjusted Pearson correlation $r_{\tilde{X}\tilde{Y}}$	0.215	0.001	0.048
Parametric adjustment (1)	10^{-4}	0.001	0.001
Partial correlation	0.040	0.001	0.003
Nonparametric partial correlation	0.040	0.001	0.003
Covariate adjusted correlation	0.014	0.001	0.001

Table 4: Estimated absolute bias, variance and MSE of the estimators of the correlation under the null case of additive parametric adjustment from proposed covariate adjusted correlation, no adjustment, parametric adjustment (1) of Tassone et al. (2000a), partial correlation, and nonparametric partial correlations under model (c) (parametric additive effects) of Section 4.1 based on 1000 Monte Carlo data sets. The results are presented for three sample sizes of $n = 150, 300$ and 600.

$n = 150$ **Estimation Method**	**Bias**	**Variance**	**MSE**
Unadjusted Pearson correlation $r_{\tilde{X}\tilde{Y}}$	0.3004	0.0009	0.0912
Parametric adjustment (1)	0.1738	0.0018	0.0320
Partial correlation	0.0020	0.0040	0.0040
Nonparametric partial correlation	0.0014	0.0042	0.0042
Covariate adjusted correlation	0.0300	0.0057	0.0065
$n = 300$			
Unadjusted Pearson correlation $r_{\tilde{X}\tilde{Y}}$	0.2994	0.0005	0.0901
Parametric adjustment (1)	0.1719	0.0009	0.0304
Partial correlation	0.0015	0.0021	0.0021
Nonparametric partial correlation	0.0022	0.0021	0.0021
Covariate adjusted correlation	0.0200	0.0026	0.0030
$n = 600$			
Unadjusted Pearson correlation $r_{\tilde{X}\tilde{Y}}$	0.3001	0.0003	0.0903
Parametric adjustment (1)	0.1737	0.0004	0.0306
Partial correlation	0.0024	0.0010	0.0010
Nonparametric partial correlation	0.0027	0.0010	0.0010
Covariate adjusted correlation	0.0149	0.0012	0.0014

Table 5: Equidistant and nearest neighbor binning for uniform, normal and χ^2 distributed U. The results are presented for three sample sizes of $n = 150, 300$ and 600.

$Unif(0.2, 0.9)$				
	Estimation Method	**Bias**	**Variance**	**MSE**
$n = 150$	Equidistant binning	0.0392	0.0060	0.0075
	Nearest neighbor binning	0.0423	0.0059	0.0077
$n = 300$	Equidistant binning	0.0223	0.0025	0.0030
	Nearest neighbor binning	0.0235	0.0026	0.0032
$n = 600$	Equidistant binning	0.0130	0.0012	0.0014
	Nearest neighbor binning	0.0132	0.0013	0.0015
$N(0.55, 0.04)$				
$n = 150$	Equidistant binning	0.0453	0.0056	0.0076
	Nearest neighbor binning	0.0510	0.0055	0.0081
$n = 300$	Equidistant binning	0.0303	0.0025	0.0034
	Nearest neighbor binning	0.0366	0.0024	0.0038
$n = 600$	Equidistant binning	0.0215	0.0011	0.0015
	Nearest neighbor binning	0.0272	0.0011	0.0018
$\chi^2(1)/6 + 0.4$				
$n = 150$	Equidistant binning	0.0335	0.0047	0.0058
	Nearest neighbor binning	0.0382	0.0050	0.0064
$n = 300$	Equidistant binning	0.0207	0.0024	0.0028
	Nearest neighbor binning	0.0204	0.0024	0.0028
$n = 600$	Equidistant binning	0.0150	0.0011	0.0013
	Nearest neighbor binning	0.0144	0.0012	0.0014

Table 6: Comparison between the proposed method and the one based on Fisher's z-transformation for uniform, normal and χ^2 distributed U. The results are presented for three sample sizes of $n = 150, 300$ and 600.

$Unif(0.2, 0.9)$				
	Estimation Method	**Bias**	**Variance**	**MSE**
$n = 150$	Proposed Method	0.0392	0.0060	0.0075
	Fisher's z	0.0423	0.0059	0.0077
$n = 300$	Proposed method	0.0223	0.0025	0.0030
	Fisher's z	0.0235	0.0026	0.0032
$n = 600$	Proposed method	0.0130	0.0012	0.0014
	Fisher's z	0.0132	0.0013	0.0015
$N(0.55, 0.04)$				
$n = 150$	Proposed method	0.0391	0.0054	0.0069
	Fisher's z	0.0196	0.0072	0.0075
$n = 300$	Proposed method	0.0333	0.0025	0.0032
	Fisher's z	0.0124	0.0030	0.0036
$n = 600$	Proposed method	0.0221	0.0013	0.0018
	Fisher's z	0.0059	0.0014	0.0015
$\chi^2(1)/6 + 0.4$				
$n = 150$	Proposed method	0.0305	0.0049	0.0058
	Fisher's z	0.0152	0.0062	0.0064
$n = 300$	Proposed method	0.0225	0.0025	0.0030
	Fisher's z	0.0025	0.0029	0.0029
$n = 600$	Proposed method	0.0169	0.0013	0.0015
	Fisher's z	0.0013	0.0014	0.0014

FIGURE CAPTIONS

1. Matrix plot of the observed variables $\widetilde{\text{CGG}}$, $\widetilde{\text{mRNA}}$ and ActRatio for $n = 165$ female premutation carriers.

2. (Top panel) Scatter plot of the local correlation estimates r_j versus U_j^M for $j = 1, \ldots, 20$ bins, with approximately 8 points per bin. A local linear smooth overlays the scatter plot with an automatically selected bandwidth of $h = 0.1$. (Bottom panel) Plot of the estimated nonparametric density (dashed line) of 1000 standardized bootstrap estimates used in forming the 95% CI's for $\rho_{\text{mRNA,CGG}}$ in the data application. The standard normal density (solid line) is also given.

3. (Top panel) Plots of $\widetilde{\rho}(u)$ from the two cases of alternatives to the proposed additive and multiplicative distortion form. The null hypothesis of additive and multiplicative forms corresponds to $\theta = 0$, i.e. the (conditional) correlation function $\widetilde{\rho}(u)$ is constant. Increasing deviation away from the null is parametrized by $\theta = 1, \ldots, 8$. (Bottom panel) Power curves for the two cases of alternatives/deviations at significance levels $\alpha = 0.05$ (bottom curve of same line type) and 0.10 for $n = 150$ (dotted curves), $n = 300$ (solid curves) and $n = 600$ (dash-dotted curves).

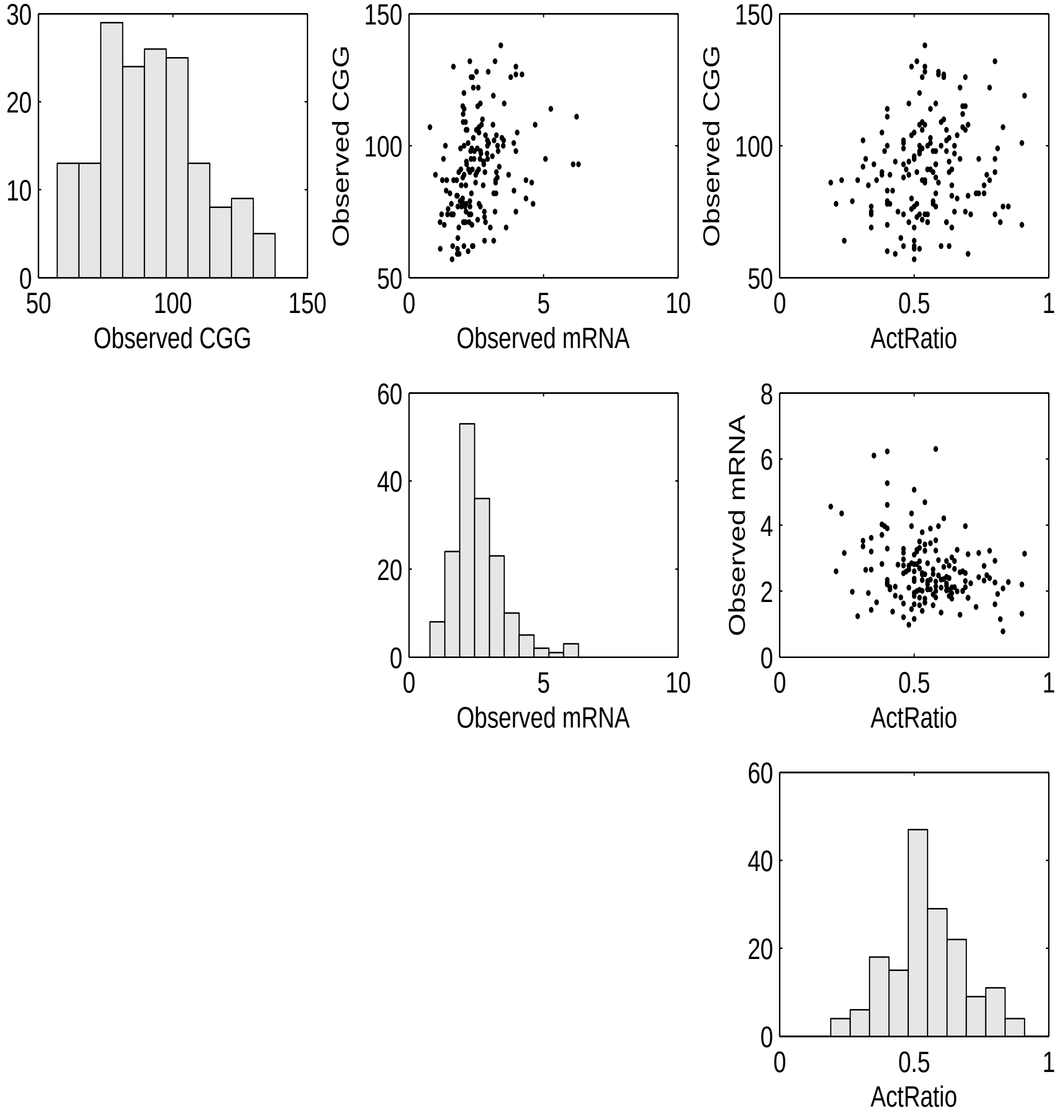
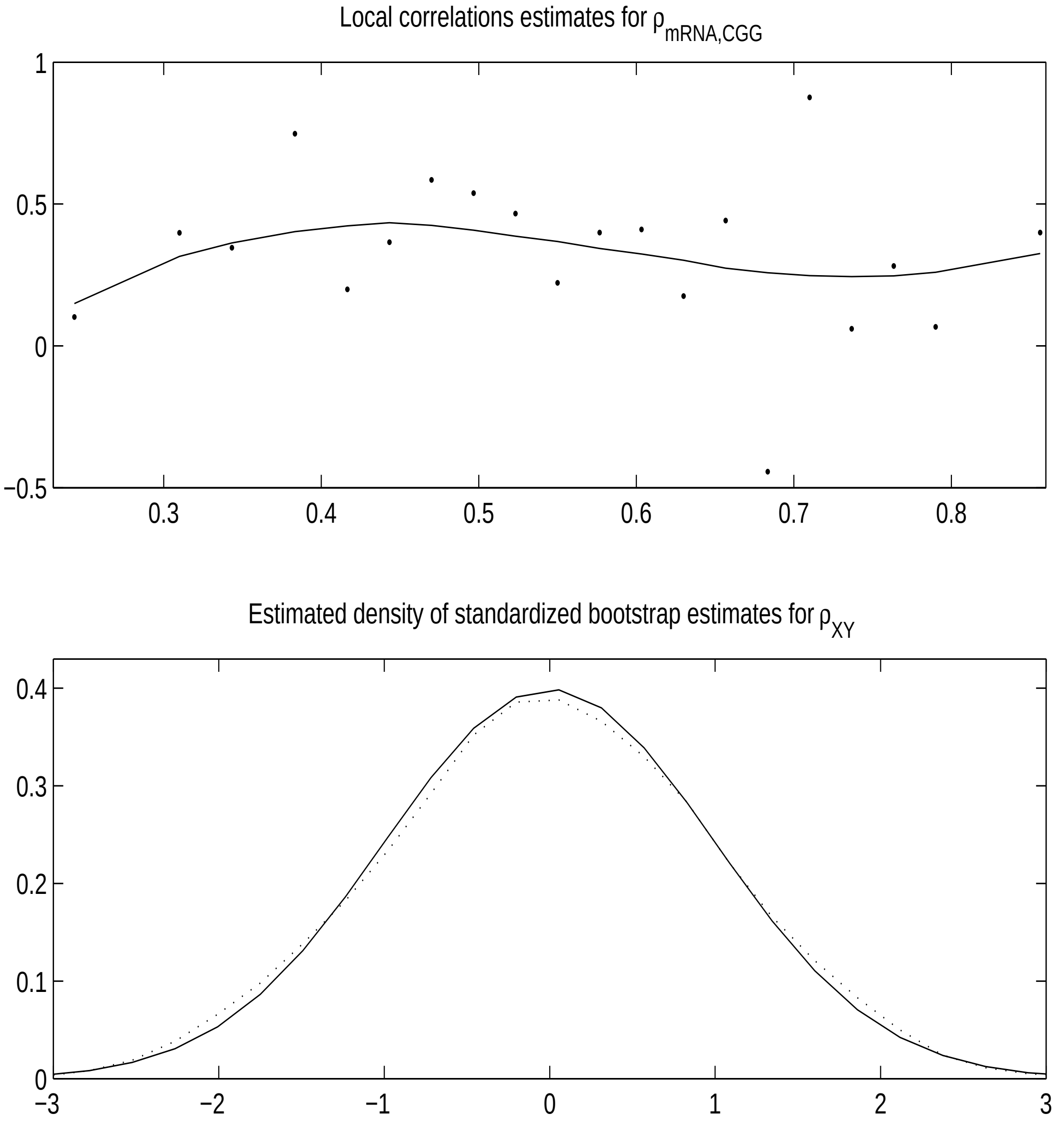

30
20
10
0
50
100
150
Observed CGG
150
100
50
0
5
10
Observed CGG
Observed mRNA
150
100
50
0
0.5
1
Observed CGG
ActRatio
60
40
20
0
0
5
10
Observed mRNA
8
6
4
2
0
0
0.5
1
Observed mRNA
ActRatio
60
40
20
0
0
0.5
1
ActRatio

Figure 1:

 Local correlations estimates for $\rho_{mRNA,CGG}$

Estimated density of standardized bootstrap estimates for ρ_{XY}

Figure 2:

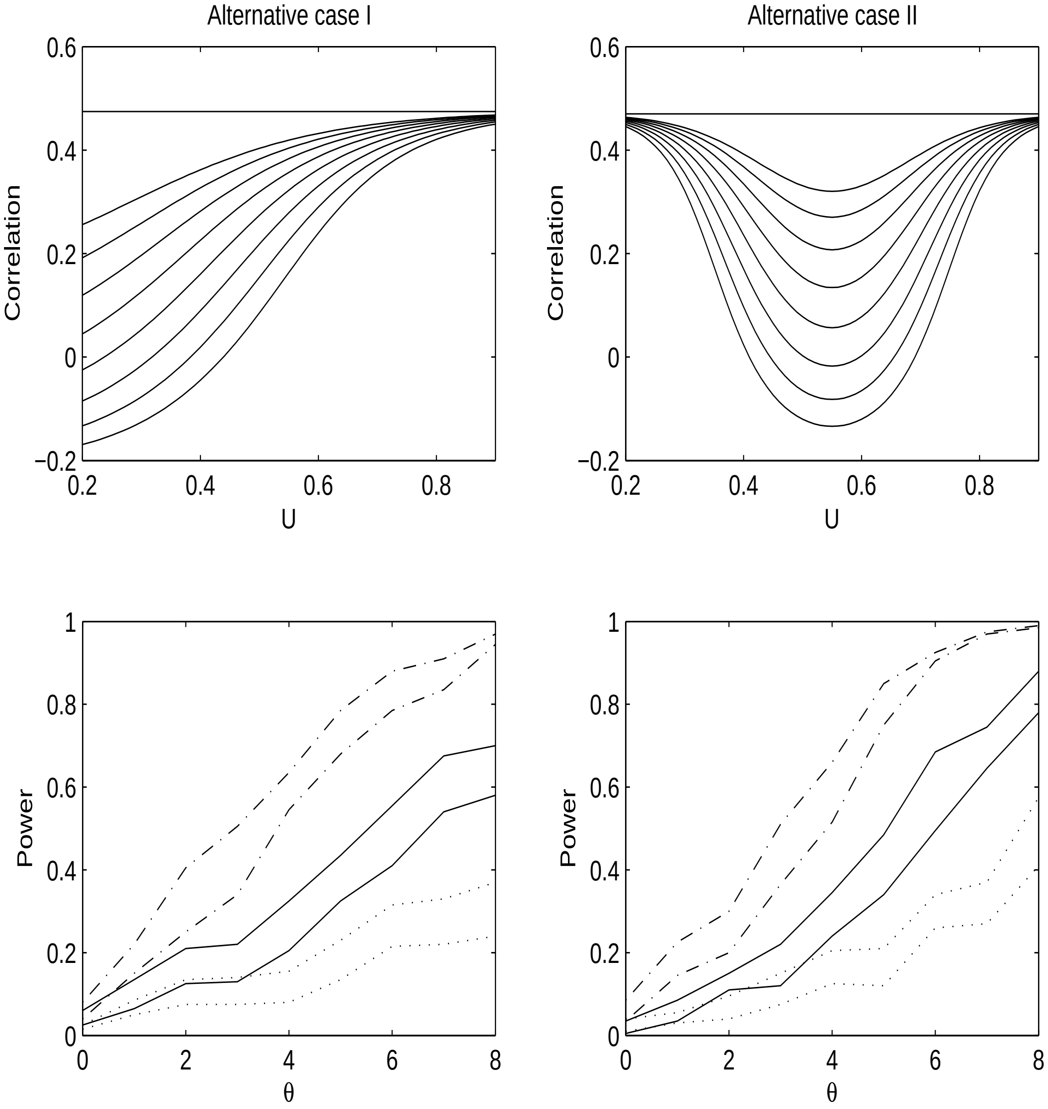

Alternative case I
Alternative case II
Correlation
Correlation
0.6
0.4
0.2
0
−0.2
0.2
0.4
0.6
0.8
U
U
Power
Power
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
θ
θ

Figure 3: