Published in final edited form as:

J Am Stat Assoc. 2014 October ; 109(508): 1533–1545. doi:10.1080/01621459.2014.908777.

A Likelihood-Based Framework for Association Analysis of Allele-Specific Copy Numbers

Y. J. Hu, D. Y. Lin, W. Sun, and D. Zeng

Abstract

Copy number variants (CNVs) and single nucleotide polymorphisms (SNPs) co-exist throughout the human genome and jointly contribute to phenotypic variations. Thus, it is desirable to consider both types of variants, as characterized by allele-specific copy numbers (ASCNs), in association studies of complex human diseases. Current SNP genotyping technologies capture the CNV and SNP information simultaneously via fluorescent intensity measurements. The common practice of calling ASCNs from the intensity measurements and then using the ASCN calls in downstream association analysis has important limitations. First, the association tests are prone to false-positive findings when differential measurement errors between cases and controls arise from differences in DNA quality or handling. Second, the uncertainties in the ASCN calls are ignored. We present a general framework for the integrated analysis of CNVs and SNPs, including the analysis of total copy numbers as a special case. Our approach combines the ASCN calling and the association analysis into a single step while allowing for differential measurement errors. We construct likelihood functions that properly account for case-control sampling and measurement errors. We establish the asymptotic properties of the maximum likelihood estimators and develop EM algorithms to implement the corresponding inference procedures. The advantages of the proposed methods over the existing ones are demonstrated through realistic simulation studies and an application to a genome-wide association study of schizophrenia. Extensions to next-generation sequencing data are discussed.

Keywords

Case-control studies; Copy number variants; Genome-wide association studies; Retrospective likelihood; Semiparametric efficiency; Single nucleotide polymorphisms

1. Introduction

SNP occurs when a single nucleotide in the DNA sequence is altered. Genome-wide association (GWA) studies have identified SNPs associated with more than 200 traits (see the National Human Genome Research Institute catalogue of published GWA studies, <http://www.genome.gov/gwastudies>). CNV refers to the amplification or deletion of a DNA segment compared to a reference genome assembly. Recent studies documented the extensive presence of CNVs in the human genome (Tuzun et al. 2005; Redon et al. 2006; McCarroll et al. 2008). Changes in the copy number can have dramatic phenotypic consequences by altering gene dosage, disrupting coding sequences or perturbing long-range gene regulation; see Lupski (2009) for a review.

Because CNVs and SNPs coexist throughout the human genome, it is desirable to consider both types of variants, as characterized by ASCNs, in association studies of complex human diseases. Ignoring CNVs during SNP genotype calling can lead to genotyping errors that appear to violate Mendelian inheritance or Hardy-Weinberg Equilibrium (HWE); therefore, SNPs in the CNV regions are typically filtered out. In addition, CNVs and SNPs may act in concert to influence disease phenotypes, as has been revealed in several cancer studies (e.g., LaFramboise et al. 2005; Van Loo et al. 2010).

Current SNP genotyping arrays can capture the ASCN information via two-dimensional measurements of fluorescent intensities. Denote the two possible alleles at a SNP site by A and B. The Affymetrix platforms represent the ASCN information by the intensities of the A and B alleles (Figure 1(a)). The Illumina platforms represent the ASCN information by Log R Ratio (LRR), which measures the total copy number, and B Allele Frequency (BAF), which measures the allelic contrast (Figure 1(b)). LRR is approximately 0 when the copy number is 2 and is positive or negative when there is copy number amplification or deletion. BAF lies between 0 and 1. BAF close to 0 or 1 implies that the locus only has the A or B allele, whereas BAF close to 0.5 implies an equal number of A and B alleles.

Several algorithms have been developed to detect/call CNVs for SNP genotyping arrays. Some of them, such as QuantiSNP (Colella et al. 2007), PennCNV (Wang et al. 2007) and GenoCNV (Sun et al. 2009), were designed specifically for Illumina data. Those algorithms rely on hidden-Markov models (HMMs) to segment the intensity measurements along the genome for each subject. PennCNV assumes the HMM parameters to be known; QuantiSNP imposes priors for the parameters so that only a few hyper-parameters need to be estimated; GenoCNV allows the parameters to be estimated from the data. GenoCNV directly calls ASCNs; PennCNV and QuantiSNP only output calls for total copy numbers, although ASCNs can be obtained by applying appropriate thresholds on the BAF measurements. For Affymetrix 6.0 arrays, a commonly used software package for CNV calling is Birdsuite (Korn et al. 2008), which handles rare and common CNVs separately.

Once the ASCN values are determined by a calling algorithm, they are commonly treated as known quantities in the downstream association analysis (Bucan et al. 2009; Diskin et al. 2009; Glessner et al. 2009; Need et al. 2009; Wang et al. 2009). This two-stage strategy, which pertains to single imputation in the missing data literature, is not optimal for two reasons. First, the association tests are not robust to differential measurement errors between cases and controls caused by differences in DNA quality or handling; see Figure 2 (upper panel) for an example of differential measurement errors. In the presence of such errors, calling ASCNs with cases and controls combined leads to differential misclassification (e.g., more cases being classified as copy number amplification than controls), which in turn results in excessive false-positive associations. Second, imputation disregards the phenotype, which may be informative about the missing data, and ignores the uncertainties in the ASCN calls. In general, imputation yields biased estimators of genetic effects and gene-environment interactions, and the variances are underestimated (Hu and Lin 2010).

Barnes et al. (2008) described a likelihood-based method for association testing with CNVs that accounts for differential measurement errors and avoids imputation. Their method has

important limitations. First, it is confined to the total copy number and ignores possible allele-specific effects. Secondly, it only makes use of the measurements for the total copy number, neglecting the fact that the two-dimensional measurements can better distinguish copy number states. Thirdly, it adopts a prospective likelihood, which may not be appropriate for case-control studies with missing data or measurement errors.

In this article, we propose a statistical framework for the integrated analysis of CNVs and SNPs in association studies, including the analysis of total copy numbers as a special case. We assume that the location and possible copy number states of each CNV have been detected by a CNV detection/calling algorithm such as PennCNV (e.g., a CNV is located at Chr1:100bp-200bp, and it may have copy numbers 0, 1, 2, and 3 in the study samples); we do not need to pre-estimate the total copy number for each individual. Our approach combines the ASCN calling and the association analysis into a single step while allowing for differential measurement errors. We formulate the effects of CNVs and SNPs on the phenotype through flexible regression models, which can accommodate various genetic mechanisms and gene-environment interactions. We focus on case-control studies, although our methods can be readily modified for other study designs and traits. We construct appropriate likelihoods, which may involve high-dimensional parameters, and we establish the consistency, asymptotic normality, and asymptotic efficiency of the maximum likelihood estimators by appealing to modern asymptotic techniques. We develop efficient and reliable numerical algorithms to implement the corresponding inference procedures. We demonstrate the advantages of the proposed methods over the existing ones through realistic simulation studies and an application to a genome-wide association study (GWAS) of schizophrenia (Shi et al. 2009). We discuss directions for future research.

2. Methods

2.1 Data and Models

Suppose that the SNP of interest has alleles A and B. Let K and L denote the total copy number and the B allele copy number, respectively, where $0 \leq L \leq K \leq N_K$, and N_K is a known integer that can be determined by a CNV detection/calling algorithm such as PennCNV. Let Y be the phenotype of interest and $\mathbf{X}$ be a set of environmental factors. For case-control studies, the conditional probability of $Y = y$ given $K = k, L = l$ and $\mathbf{X} = \mathbf{x}$ is formulated through the logistic regression model

$$P_{\alpha, \beta}(y|k, l, \mathbf{x}) = \frac{\exp \{y(\alpha + \beta^T \mathbf{Z}(k, l, \mathbf{x}))\}}{1 + \exp \{\alpha + \beta^T \mathbf{Z}(k, l, \mathbf{x})\}}, \quad (1)$$

where $\mathbf{Z}(k, l, \mathbf{x})$ is a design vector excluding the unit component. The vector $\{1, \mathbf{Z}(K, L, \mathbf{X})^T\}$ is assumed to be linearly independent. There is considerable flexibility in specifying this model. A linear predictor in the form of $\alpha + \beta k$ pertains to an additive model for the total copy number and $\alpha + \beta_1 I(k=1) + \dots + \beta_{N_K} I(k=N_K)$ to a saturated model, where $I(\cdot)$ is the indicator function. Replacing k in the linear predictors by l yields the additive and saturated models for the B allele copy number. We may also specify the linear predictor as $\alpha + \beta_1 k$

$+\beta_2\{(k-l)-l\}$, in which case β_1 and β_2 correspond to the effects of the total copy number and allelic difference, respectively.

Although we are interested in the effects of $(K, L, \mathbf{X})$ on Y , we observe intensity measurements, denoted by $\mathbf{R}$, instead of (K, L) . Thus, we have a regression problem with measurement errors. We describe below how to model the measurement distribution $P(\mathbf{R}|Y, K, L, \mathbf{X})$. We allow the distribution to depend on the disease group and environmental factors in addition to the ASCN states so as to account for differential measurement errors. The specific form of $P(\mathbf{R}|Y, K, L, \mathbf{X})$ depends on the type of SNP array.

2.1.1 Affymetrix Data—Each SNP has a pair of measurements, denoted by (R_A, R_B) , for the A allele and B allele intensities, respectively. We assume that (R_A, R_B) given $\mathbf{C} \equiv (Y, K, L, \mathbf{X})$ follows a bivariate normal distribution

$$P_{\mu, \Sigma}(R_A, R_B|\mathbf{C}) = \phi \left\{ \begin{bmatrix} R_A \\ R_B \end{bmatrix}; \mu_{\mathbf{C}}, \Sigma_{\mathbf{C}} \right\},$$

where $\phi(\cdot; \mu_{\mathbf{C}}, \Sigma_{\mathbf{C}})$ is the bivariate normal density function with mean $\mu_{\mathbf{C}}$ and covariance matrix $\Sigma_{\mathbf{C}}$, $\mu = \{\mu_{\mathbf{C}}\}$ and $\Sigma = \{\Sigma_{\mathbf{C}}\}$. We specify the mean and covariance matrix given $\mathbf{C}$ by a saturated model of five parameters and allow the parameters to vary among SNP sites. If certain covariates such as age are assumed to be independent of (R_A, R_B) , we drop them from $\mathbf{C}$. Continuous components of $\mathbf{X}$ are assumed to be independent of (R_A, R_B) or be discretized.

2.1.2 Illumina Data—In Illumina platforms, R_A and R_B are transformed to measures of LRR and BAF, which are denoted as R_{LRR} and R_{BAF} , respectively; see the Methods section of Wang et al. (2007) for details. The Illumina transformation adjusts for different chemical characteristics of each SNP so that the heterogeneity of intensity measurements across SNPs is reduced. By their definitions, R_{LRR} and R_{BAF} can be treated as independent given $(Y, K, L, \mathbf{X})$. We model the conditional distribution of R_{LRR} given $\tilde{\mathbf{C}} \equiv (Y, K, \mathbf{X})$ by a normal density function

$$\phi(R_{\text{LRR}}; \mu_{\text{LRR}, \tilde{\mathbf{C}}}, \sigma_{\text{LRR}, \tilde{\mathbf{C}}}^2), \quad (2)$$

where $\phi(\cdot; \mu_{\text{LRR}, \tilde{\mathbf{C}}}, \sigma_{\text{LRR}, \tilde{\mathbf{C}}}^2)$ is the univariate normal density function with mean $\mu_{\text{LRR}, \tilde{\mathbf{C}}}$ and variance $\sigma_{\text{LRR}, \tilde{\mathbf{C}}}^2$, which are parameters specific to each $\tilde{\mathbf{C}}$. Note that the conditional distribution of R_{LRR} does not depend on L . By definition, R_{BAF} should be around 0, 0.5 and 1 for genotypes AA, AB and BB, respectively, and truncated at 0 and 1. Deviation of the R_{BAF} value from these three values may indicate a CNV. For instance, a R_{BAF} value of 0.33 may indicate genotype AAB. We formulate the conditional distribution of R_{BAF} given $\mathbf{C}$ by a truncated normal density function

$$\phi(R_{\text{BAF}}; \mu_{\text{BAF},C}, \sigma_{\text{BAF},C}^2)^{I(0 < R_{\text{BAF}} < 1)} \Phi(0; \mu_{\text{BAF},C}, \sigma_{\text{BAF},C}^2)^{I(R_{\text{BAF}}=0)} \times (1 - \Phi(1; \mu_{\text{BAF},C}, \sigma_{\text{BAF},C}^2))^{I(R_{\text{BAF}}=1)}, \quad (3)$$

where $\Phi(\cdot; \mu_{\text{BAF},C}, \sigma_{\text{BAF},C}^2)$ is the distribution function corresponding to $\phi(\cdot; \mu_{\text{BAF},C}, \sigma_{\text{BAF},C}^2)$. Parameter estimation in the presence of the truncated normal distribution is numerically challenging. In Appendix A.1, we describe a strategy to avoid the truncation. In particular, we assume the means to be 0 and 1 for homozygous genotypes except for $K = 0$, so that we only need to estimate the variances. We write $\boldsymbol{\mu} = \{\mu_{\text{LRR},C}, \mu_{\text{BAF},C}\}$ and $\boldsymbol{\Sigma} = \{\sigma_{\text{LRR},C}^2, \sigma_{\text{BAF},C}^2\}$ and use $P_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(R_{\text{LRR}}, R_{\text{BAF}} | Y, K, L, \mathbf{X})$ to denote the product of (2) and (3). Clearly, the measurement model for Illumina data involves fewer components in $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ than that of Affymetrix data, which can lead to more efficient association analysis.

2.2 Association Analysis of ASCNs

Write $\mathbf{R} = (R_A, R_B)$ for Affymetrix data and $\mathbf{R} = (R_{\text{LRR}}, R_{\text{BAF}})$ for Illumina data. Let $F(\cdot)$ and $f(\cdot)$ be the distribution and density functions of $\mathbf{X}$. In some applications, $\mathbf{X}$ and (K, L) are correlated. One important example is when $\mathbf{X}$ represents the principal components for ancestry. Another example is when a gene influences both environmental exposure (e.g., cigarette smoking) and disease occurrence (e.g., lung cancer) (Amos et al. 2008). In such cases, we allow gene-environment dependence by leaving the probability of $(K = k, L = l)$ given $\mathbf{X} = \mathbf{x}$, denoted by $G(k, l | \mathbf{x})$, completely unspecified. To account for the case-control sampling, we adopt the retrospective likelihood $\prod_{i=1}^n P(\mathbf{R}_i, \mathbf{X}_i | Y_i)$, which takes the form

$$L_r(\boldsymbol{\theta}, G, F) = \prod_{i=1}^n \left\{ \frac{\sum_{k=0}^{N_K} \sum_{l=0}^k P_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\mathbf{R}_i | Y_i, k, l, \mathbf{X}_i) P_{\alpha, \beta}(Y_i | k, l, \mathbf{X}_i) G(k, l | \mathbf{X}_i) f(\mathbf{X}_i)}{\sum_{k=0}^{N_K} \sum_{l=0}^k \int_{\mathbf{x}} P_{\alpha, \beta}(Y_i | k, l, \mathbf{x}) G(k, l | \mathbf{x}) dF(\mathbf{x})} \right\},$$

where $\boldsymbol{\theta} = (\alpha, \beta, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, and n is the number of study subjects. Note that the distribution of the observed data $(\mathbf{R}_i, Y_i, \mathbf{X}_i)$ for the i th subject is modeled by a mixture of bivariate normal

distributions. By contrast, the prospective likelihood $\prod_{i=1}^n P(\mathbf{R}_i, Y_i, \mathbf{X}_i)$ takes the form

$$L_p(\boldsymbol{\theta}, G) = \prod_{i=1}^n \left\{ \sum_{k=0}^{N_K} \sum_{l=0}^k P_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\mathbf{R}_i | Y_i, k, l, \mathbf{X}_i) P_{\alpha, \beta}(Y_i | k, l, \mathbf{X}_i) G(k, l | \mathbf{X}_i) \right\}. \quad (4)$$

For the logistic regression analysis of standard case-control data, Prentice and Pyke (1979) established that the retrospective and prospective likelihood functions yield the same estimators of odds ratios provided that the distribution of covariates (genetic and environmental factors) is completely unspecified. Roeder et al. (1996) showed that the equivalence continues to hold when covariates are measured with errors. By the arguments of Roeder et al. (1996), the maximum likelihood estimator of $(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ based on the

retrospective likelihood $L_r(\theta, G, F)$ can be obtained by maximizing the prospective likelihood $L_p(\theta, G)$ when the distribution of $(K, L, \mathbf{X})$ is completely unspecified.

Because K and L are not observable, the estimation of $G(k, l|\mathbf{x})$ requires the distribution of (K, L) conditional on $\mathbf{X}$ to change smoothly with $\mathbf{X}$. We use the sieve MLE approach (Shen

1997), under which $G(k, l|\mathbf{x})$ is approximated by $\sum_{m=1}^{M_n} g_m(k, l) B_m(\mathbf{x})$ in the maximization, where M_n depends on n , the $g_m(k, l)$'s are probability mass functions, and the $B_m(\mathbf{x})$'s are tensor products of B-splines. When $\mathbf{X}$ is one-dimensional, it suffices to use a simple form of B-splines, i.e., a piece-wise constant function, which leads to the histogram sieve, i.e.,

$G(k, l|x) = \sum_{m=1}^{M_n} g_m(k, l) I(x \in \chi_m)$, where $\chi_1, \dots, \chi_{M_n}$ are equally spaced bins of the X domain. The maximization can be carried out by the EM algorithm described in Appendix A.2. The resulting sieve MLEs are denoted by $\hat{\theta}$ and $\hat{G}$, whose asymptotic properties are stated in Theorem 1. For any parameter χ , we denote its true value by χ_0 when the distinction is necessary. We assume that the true value of any Euclidean parameter χ belongs to the interior of a known compact set within the domain of χ .

THEOREM 1. Under mild regularity conditions, $\hat{\theta} \xrightarrow{p} \theta_0$ almost surely, and $n^{1/2}(\hat{\theta} - \theta_0)$ converges in distribution to a zero-mean normal random vector whose covariance matrix attains the semiparametric efficiency bound.

The proofs of this theorem and other theorems are provided in Appendix B.

In many applications, it is appropriate to assume gene-environment independence, so that $G(K, L|\mathbf{X})$ does not depend on $\mathbf{X}$. Under HWE, L given K is binomial with success probability p_B , which is the population frequency of the B allele. We denote the binomial distribution by $P_{p_B}(L|K)$ and write $\pi_k = P(K = k)$, so that $P(K = k, L = l) = P_{p_B}(l|k) \pi_k$. If we impose such structures in the covariate distribution, the equivalence between the retrospective and prospective likelihoods no longer holds and the retrospective likelihood should be used. There is very little information about α in the retrospective likelihood, so the problem is virtually nonidentifiable. One possible solution is to assume that the disease is rare, so that we can approximate (1) by $P_{\alpha, \beta}(y|k, l, \mathbf{x}) \approx \exp \left\{ y \left(\alpha + \beta^T Z(k, l, \mathbf{x}) \right) \right\}$. Let $\theta = (\beta, \pi, p_B, \mu, \Sigma)$. The retrospective likelihood can then be approximated by

$$\tilde{L}_r(\theta, F) = \prod_{i=1}^n \left[\frac{\sum_{k=0}^{N_K} \sum_{l=0}^k P_{\mu, \Sigma}(\mathbf{R}_i | Y_i, k, l, \mathbf{X}_i) \exp \left\{ Y_i \beta^T Z(k, l, \mathbf{X}_i) \right\} P_{p_B}(l|k) \pi_k f(\mathbf{X}_i)}{\sum_{k=0}^{N_K} \sum_{l=0}^k \int_{\mathbf{x}} \exp \left\{ Y_i \beta^T Z(k, l, \mathbf{x}) \right\} P_{p_B}(l|k) \pi_k dF(\mathbf{x})} \right], \quad (5)$$

in which α disappears. We use the nonparametric maximum likelihood estimation (NPMLE), in which the distribution function F is treated as a right-continuous function with jumps at the observed $\mathbf{X}$ and the objective function to be maximized is obtained from $L_r(\theta, F)$ by replacing $f(\mathbf{x})$ with the jump size of F at $\mathbf{x}$. The maximization can be carried out by the EM algorithm of Appendix A.3. The resulting NPMLEs are denoted by $\hat{\theta}$ and $\hat{F}$, whose asymptotic properties are stated in Theorem 2.

THEOREM 2. Under mild regularity conditions, $|\hat{\theta} - \theta_0| + \sup_{\mathbf{x}} |F(\hat{\mathbf{x}}) - F_0(\mathbf{x})| \rightarrow 0$ almost surely, and $n^{1/2}(\hat{\theta} - \theta_0)$ converges in distribution to a zero-mean normal random vector whose covariance matrix attains the semiparametric efficiency bound.

In Theorems 1-2, the limiting covariance matrix of $\hat{\theta}$ can be consistently estimated by the Louis (1982) formula or by the profile likelihood method (Murphy and van der Vaart 2000). When the disease is not rare but the disease rate is known, the likelihood can be constructed along the lines of Hu, Lin and Zeng (2010). In the sequel, all hypothesis tests are based on the Wald statistics.

2.3 Total Copy Number

In some cases, such as the data from the copy number probes of Affymetrix 6.0 arrays or the array comparative genomic hybridization (CGH), only the total copy number is measured at a locus. Let R be the one-dimensional measurement of the total copy number and $P_{\mu, \Sigma}(R|Y, K, \mathbf{X})$ be a univariate normal density function. We can readily accommodate such cases by reducing (4) and (5) to

$$L_p(\theta, G) = \prod_{i=1}^n \left\{ \sum_{k=0}^{N_K} P_{\mu, \Sigma}(R_i Y_i, k, \mathbf{X}_i) P_{\alpha, \beta}(Y_i | k, \mathbf{X}_i) G(k | \mathbf{X}_i) \right\}, \quad (6)$$

where $\theta = (\alpha, \beta, \mu, \Sigma)$, and

$$\tilde{L}_r(\theta, F) = \prod_{i=1}^n \left[\frac{\sum_{k=0}^{N_K} P_{\mu, \Sigma}(R_i Y_i, k, \mathbf{X}_i) \exp \left\{ Y_i \beta^T Z(k, \mathbf{X}_i) \right\} \pi_k f(\mathbf{X}_i)}{\sum_{k=0}^{N_K} \int_{\mathbf{x}} \exp \left\{ Y_i \beta^T Z(k, \mathbf{x}) \right\} \pi_k dF(\mathbf{x})} \right], \quad (7)$$

where $\theta = (\beta, \pi, \mu, \Sigma)$. Barnes et al. (2008) dealt with this problem by adopting a prospective likelihood

$$\prod_{i=1}^n \left\{ \sum_{k=0}^{N_K} P_{\mu, \Sigma}(R_i Y_i, k, \mathbf{X}_i) P_{\alpha, \beta}(Y_i | k, \mathbf{X}_i) P_{\zeta}(k | \mathbf{X}_i) \right\}, \quad (8)$$

where $P_{\zeta}(k|\mathbf{x})$ is a multinomial regression model of $K = k$ given $\mathbf{X} = \mathbf{x}$ with parameters ζ . When there are no environmental factors, (8) is identical to (6); therefore, Barnes et al.'s method is justified. In the presence of environmental factors and gene-environment dependence, Barnes et al. (2008) imposed a parametric structure on $P(K|\mathbf{X})$ whereas we leave the distribution to be nonparametric. The nonparametric nature of $P(K|\mathbf{X})$ is the key to the use of the prospective likelihood under case-control sampling. With any parametric constraint, the prospective likelihood is no longer appropriate and the retrospective likelihood should be used instead.

3. Simulation Studies

We conducted extensive simulation studies to evaluate the performance of the proposed and existing methods in realistic settings. We considered primarily the case of CNV deletion, in which K takes values 0, 1 and 2 with probabilities 0.05, 0.35 and 0.6, and given K , the B allele copy number L follows a binomial distribution with success probability 0.6. The parameter values were estimated from the data on SNP rs983985 in a GWAS study of schizophrenia (Shi et al. 2009); see Figure 1(b). We set α in model (1) to -4.6 to yield disease rates of about 1%, and obtained 1,000 controls and 1,000 cases. For controls, we generated intensity measurements from (truncated) normal distributions with the means and variances estimated at the SNP in Figure 1(b). The intensity measurements of cases were allowed to have different means or variances from controls. We obtained 5,000 replicates.

In the first set of studies, we evaluated the performance of the proposed methods based on the prospective likelihood (4) with a completely unspecified distribution of the covariates (K , L , X) and on the retrospective likelihood (5) with the assumption of gene-environment independence (i.e., independence between genetic and environmental factors). We considered the disease model with a gene-environment interaction: $\text{logit}P(Y = 1|K, L, X) = \alpha + \beta_1 K + \beta_2 X + \beta_3 KX$, where $X \sim N(cK, 1)$, and c was set to 0 and 0.1 to reflect gene-environment independence and dependence, respectively. The results for studying the gene-environment interaction are summarized in Table 1. In the presence of gene-environment dependence, likelihood (4) works well whereas likelihood (5) yields large bias and inflated type I error. When the independence assumption holds, (5) can gain substantial efficiency over (4).

Our second set of studies was designed to assess the sensitivity of the type I error to the differences of means or variances of intensity measurements between cases and controls. We considered the disease model with an additive effect of the B allele copy number: $\text{logit}P(Y = 1|K, L) = \alpha + \beta_0 L$. We applied our method based on (5) as well as two imputation methods. The two imputation methods use a two-dimensional Gaussian mixture model (GMM) to assign each subject the most likely ASCN state, one fitting the GMM with cases and controls combined and one fitting it for cases and controls separately; they are referred to as imputation-C and imputation-S, respectively. As shown in Figure 3 (a), the type I error of imputation-C increases rapidly as the mean of the ASCN measurements in cases shifts away from that of controls. The type I error of Imputation-S remains constant with respect to the mean shift but is always inflated; the inflation results from over-estimating the differences in the ASCN frequency between cases and controls because nuisance parameters are allowed to vary between the two groups and from ignoring the uncertainties in the ASCN assignment. Figure 3 (b) shows that the type I error inflation of imputation-S grows as the variance of the measurement in cases increases. Imputation-C is robust to differential variances in that it does not generate differential misclassification. The proposed method provides the most robust test by modeling the measurements of cases and controls separately and accounting for all uncertainties.

CNV amplification has lower signal-to-noise ratio and is more difficult to impute. In the third set of simulation studies, we modified the setting of the second set by replacing the

deletion allele with a duplication one and mimicking the intensity distribution seen in the lower panel of Figure 2. As shown in Figure 3 (c) and (d), the inflation of the type I error for the two imputation methods are more profound than in (a) and (b). The proposed method continues to have correct control of the type I error.

In the fourth set of studies, we assumed no differences in the measurement mean or variance between cases and controls in order to separate the influence of imputation itself from that of differential measurement errors. The disease model was the same as in the second set of studies. Table 2 displays the results for various values of β_0 . The estimator of β_0 produced by imputation-C is biased towards the null. The reason is that ignoring the phenotype when inferring the ASCN states makes the imputed ASCN states more homogeneous between cases and controls than they really are. As a result, imputation-C has lower power than the proposed method. Under the null, imputation-C yields an unbiased estimator of β_0 , a correct variance estimator, and accurate type I error; see Hu and Lin (2010) for a proof in the context of SNP association analysis. As expected, the variance estimator of imputation-S underestimates the true variance. The discrepancy remains under the null, causing inflated type I error. Consequently, imputation-S can be more powerful than the proposed method. The proposed method yields unbiased effect estimator, correct variance estimator, accurate type I error and reasonable power.

Barnes et al. (2008)'s method can be used to analyze the total copy number. While our method makes use of both the LRR and BAF measurements, Barnes et al.'s method only uses the LRR measurements. We compared the two methods in the fifth set of simulation studies, in which the disease status was generated from the logistic regression model: $\text{logit}P(Y = 1|K) = \alpha + \beta_0 K$. The results are summarized in Table 3. Both methods yield unbiased estimators of β_0 with correct variance estimators and thus accurate type I error. However, the proposed method is substantially more powerful; see Figure 4(a).

It is also of interest to compare the proposed and Barnes et al.'s methods when only the measurements on the total copy number are available. Because the two methods are equivalent in the absence of environmental factors, we focused on testing gene-environment interactions in the last set of studies. We adopted the disease model: $\text{logit}P(Y = 1|K, X) = \alpha + \beta_1 K + \beta_2 X + \beta_3 KX$, where X is independent of K . We generated the LRR measurements mimicking SNP rs983985 in the schizophrenia data; see Figure 1(c). Because of the gene-environment independence, we adopted the retrospective likelihood given in (7). As shown in Figure 4(b), the proposed method is substantially more powerful than Barnes et al.'s method because our retrospective likelihood exploits gene-environment independence whereas Barnes et al.'s prospective likelihood does not.

4. Schizophrenia Data

Schizophrenia is a severe psychiatric disorder with a lifetime prevalence of 0.4-1%. It is highly heritable and genetically heterogeneous. Recent studies showed that common SNPs and rare, large CNVs are associated with schizophrenia (Shi et al. 2009; Stefansson et al. 2008, 2009; The International Schizophrenia Consortium 2008, 2009), but common CNVs and the joint effects of common CNVs and common SNPs have not been investigated. Here,

we assess the associations of common ASCNs with schizophrenia using the case-control GWAS data of the Molecular Genetics of Schizophrenia (MGS) (Shi et al. 2009). MGS collected unrelated adult cases with DSM-IV schizophrenia from ten sites in the United States and Australia and recruited unrelated adult controls through Knowledge Networks by phone calls. Part of the MGS sample was genotyped by the Affymetrix 6.0 platform at the Broad Institute under the support of the Genetic Association Information Network (GAIN) and is thus referred to as the GAIN sample. The GAIN sample consists of both European ancestry (EA) and African American (AA) subjects and our analysis pertains only to the EA portion of the GAIN sample, which includes 1172 cases and 1378 controls. The different collection processes of cases and controls imply the possibility of differential measurement errors. Indeed, if we treat all controls as if they were from the 11th site, the principal components calculated from the intensity measurements are correlated with many of these sites (results not shown). We describe the pre-processing of the data in Appendix A.4.

Affymetrix 6.0 arrays contain more than 906,600 SNP probes and more than 946,000 copy number probes. At each SNP probe, we consider the disease model:

$$\text{logit } P(Y=1|K, L) = \alpha + \beta_1 K + \beta_2 \{(K - L) - L\} + \beta_3 \text{ gender} + \beta_4 \text{ age}.$$

A two-degrees-of-freedom test of the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ provides a joint test of the copy number and allelic effects; $H_0 : \beta_1 = 0$ pertains to a test of the copy number effect controlling for the allelic variation; $H_0 : \beta_2 = 0$ corresponds to a test of the allelic effect controlling for the copy number variation. As shown in Figure 5, the quantile-quantile (Q-Q) plot for the proposed method based on (5) agrees well with the global null distribution whereas those of imputation-C and imputation-S deviate substantially from the global null distribution. These results illustrate that the proposed method has correct type I error while imputation-C and imputation-S have inflated type I error. The inflation of imputation-C is expected because 81 out of 250 SNPs have the values of $|\mu|$ greater than 0.05. The results for the SNP displayed in the upper panel of Figure 2 are shown in the upper panel of Table 4. This SNP corresponds to the top hit of imputation-C. Consistent with the observation in the upper panel of Figure 2 that there are serious differential measurement errors in the intensity measurements but no appreciable differences of ASCN frequencies between cases and controls, the proposed method yields non-significant p-values for all three tests. The significant association identified by imputation-C is likely due to its sensitivity to differential means, and the significant association identified by imputation-S is likely due to its sensitivity to measurement variances, especially when the variances are so large that different clusters are not well separated. The results for the SNP displayed in the middle panel of Figure 2 are shown in the middle panel of Table 4. This SNP corresponds to the top hit of the proposed method. Because the middle panel of Figure 2 shows negligible differential measurement error but some differences in the ASCN frequency, the proposed and imputation methods produce comparably low p-values. The results for the duplicated SNP displayed in the lower panel of Figure 2 are shown in the lower panel of Table 4. The results of imputation-C and imputation-S may not be reliable because Figure 2 suggests that there is high uncertainty in assigning the value of 2 or 3 to the total copy number.

5. Discussion

The proposed framework is very general in several aspects: (1) it provides the integrated analysis of CNVs and SNPs as well as the analysis of CNVs only; (2) it formulates the effects of CNVs and SNPs on the phenotype through flexible regression models, which can accommodate various genetic mechanisms and gene-environment interactions; (3) it allows genetic and environmental variables to be correlated while leaving the distribution of environmental variables completely unspecified; (4) it can be readily extended to other study designs and traits; and (5) it can accommodate both Affymetrix and Illumina data, as well as all platforms that assay CNVs quantitatively, such as array CGH.

The program that implements the proposed methods is very fast and scalable to genome-wide association scans. It took about 2 hrs on a 64-bit, 3.0-GHz Intel Xeon machine to perform the analysis on chromosome 1 of the schizophrenia data. The relevant software, named CNVstat, is posted at <http://dlin.web.unc.edu/software/cnvstat/>.

The imputation approach is popular in the association analysis of CNVs. Hu and Lin (2010) showed that the imputation approach has proper control of type I error for single-SNP analysis provided that the imputation does not depend on the phenotype. This is also true of the CNV association analysis in the absence of differential measurement errors. However, differential measurement errors are prevalent and difficult to avoid, as case and control samples can rarely be obtained in strictly comparable circumstances to ensure identical DNA handling. Therefore, the imputation approach is not recommended for the association analysis of CNVs.

We only deal with common CNVs. Our methods rely on prior information of common CNVs (e.g. their locations and copy number states), which can be obtained by running a CNV calling algorithm or by appealing to a reference map of common CNVs. Thus, our approach is a complement, rather than a competitor, to PennCNV, QuantiSNP and GenoCNV. Our algorithm works well when the frequency of the deletion or duplication allele is greater than 3% with a sample size of 1,000 cases and 1,000 controls.

We do not pool information across SNPs in parameter estimation for several reasons. First, the intensity at a locus depends on the restriction fragment length and the GC content, so there is no simple way of pooling. Secondly, because CNV calling methods such as PennCNV tend to give different boundaries of CNV regions for different subjects, it is difficult to determine a common boundary. Thirdly, although the total copy numbers may be the same among SNPs in a CNV region, the allele-specific copy numbers may not.

Bayesian information criterion can be used to conduct model selection. We may select the number of copy number states to reflect the presence of deletion, amplification or both, although the number was assumed known from PennCNV in the analysis of the schizophrenia data. We may replace the saturated model for the measurement distribution by a constrained model. We may use prior information to determine whether differential measurement errors exist.

High-throughput DNA sequencing provides comprehensive CNV and SNP measurements, and several methods have been proposed to call CNVs from sequencing data (e.g., Medvedev et al. 2009; Mills et al. 2011; Handsaker et al. 2011). Read depth (i.e., the number of reads that overlap with the SNP or CNV locus), split reads (i.e., a single read that is mapped to at least two disjoint genomic regions) and read pair length (i.e., the length of a paired-end read after mapping both ends to a reference genome) are all informative about copy number changes. We are currently extending our framework to incorporate such information.

Appendix A: Numerical Algorithms

A.1 Strategy to avoid the truncated normal distribution

Because R_{BAF} measures the proportion of the signal intensity of the B allele out of the total intensity, we assume the means to be 0 or 1 for the homozygous genotypes except for $K = 0$, so we only need to estimate the variances. Specifically, we use the observed R_{BAF} such that $0 < R_{\text{BAF}} < 1$ to estimate the variances for the homozygous genotypes except for $K = 0$. For $K = 0$ or heterozygous genotypes when $K > 0$, we assume that the mean values are away from the boundary (0 or 1) and the variances are sufficiently small so that the probability of being truncated is negligible.

A.2 EM Algorithm to maximize (4)

Here, we present the EM algorithm for the histogram sieve, in which K and L are missing variables. The algorithm for the general B-spline sieve is similar.

The complete-data log-likelihood function pertaining to (4) is

$$\sum_{i=1}^n \sum_{m=1}^{M_n} I(X_i \in \chi_m) \sum_{k,l} I(K_i=k, L_i=l) \log \left\{ P_{\mu, \Sigma}(\mathbf{R}_i | Y_i, k, l, X_i) P_{\alpha, \beta}(Y_i | k, l, X_i) g_m(k, l) \right\}.$$

In the E-step, we evaluate $P(K_i = k, L_i = l | \mathbf{R}_i, Y_i, X_i)$, which can be shown to be

$$\omega_{ikl} \equiv \sum_{m=1}^{M_n} I(X_i \in \chi_m) \frac{P_{\mu, \Sigma}(\mathbf{R}_i | Y_i, k, l, X_i) P_{\alpha, \beta}(Y_i | k, l, X_i) g_m(k, l)}{\sum_{k', l'} P_{\mu, \Sigma}(\mathbf{R}_i | Y_i, k', l', X_i) P_{\alpha, \beta}(Y_i | k', l', X_i) g_m(k', l')}.$$

In the M-step, we calculate $g_m(k, l) = \sum_{i=1}^n \omega_{ikl} I(X_i \in \chi_m) / \sum_{i=1}^n I(X_i \in \chi_m)$ and

maximize $\sum_{i,k,l} \omega_{ikl} \log \left\{ P_{\mu, \Sigma}(\mathbf{R}_i | Y_i, k, l, X_i) P_{\alpha, \beta}(Y_i | k, l, X_i) \right\}$ via the Newton-Raphson algorithm. The starting values of the parameters are as follows: set $\alpha = 0$, $\beta = 0$, $g_m(k, l)$ to be discrete uniform density functions, and (μ, Σ) to be the empirical means and variances of the clusters classified by a CNV calling method. Starting with such values, we iterate until the change of the observed log-likelihood is negligible.

A.3 EM Algorithm to maximize (5)

When there are environmental factors $\mathbf{X}$, we first profile out their distribution function F from (5). Let n_1 and n_0 be the numbers of cases and controls, respectively. Suppose that there are J distinct observed values of $\mathbf{X}$, denoted by $\mathbf{x}_1, \dots, \mathbf{x}_J$. Let n_{+j} be the number of times that $\mathbf{x}_j$ is observed in the data and let η_j be the jump size of F at $\mathbf{x}_j$. Note that

$\sum_{j=1}^J \eta_j = 1$. We show that $\boldsymbol{\eta} = (\eta_1, \dots, \eta_J)^T$ can be profiled out from (5) by introducing one free parameter ν . The logarithm of (5) is

$$\begin{aligned} \tilde{l}_r(\boldsymbol{\theta}, \boldsymbol{\eta}) \\ = \sum_{i=1}^n \log \left\{ \sum_{k,l} P_{\mu, \sum}(\mathbf{R}_i Y_i, k, l, \mathbf{X}_i) e^{Y_i \beta^T \mathbf{Z}(k, l, \mathbf{X}_i)} P_{pB}(\boldsymbol{\ell}_k) \pi_k \right\} + \sum_{j=1}^J n_{+j} \log \eta_j - n_1 \log \left\{ \sum_{j,k,l} e^{\beta^T \mathbf{Z}(k, l, \mathbf{x}_j)} P_{pB}(\boldsymbol{\ell}_k) \pi_k \eta_j \right\}. \end{aligned}$$

We introduce a Lagrange multiplier λ for $\sum_{j=1}^J \eta_j = 1$ and set the derivative with respect to η_j to zero:

$$n_{+j} \eta_j^{-1} - \left\{ n_1 \sum_{k,l} e^{\beta^T \mathbf{Z}(k, l, \mathbf{x}_j)} P_{pB}(\boldsymbol{\ell}_k) \pi_k \right\} \left\{ \sum_{j,k,l} e^{\beta^T \mathbf{Z}(k, l, \mathbf{x}_j)} P_{pB}(\boldsymbol{\ell}_k) \pi_k \eta_j \right\}^{-1} - \lambda = 0.$$

Multiplying both sides by η_j and summing over $j = 1, \dots, J$, we have $\lambda = n_0$. Thus,

$$\eta_j = n_{+j} \left\{ n_0 + n_1 \sum_{k,l} e^{\beta^T \mathbf{Z}(k, l, \mathbf{x}_j)} P_{pB}(\boldsymbol{\ell}_k) \pi_k / \nu \right\}^{-1}, \text{ where } \nu = \sum_{j,k,l} e^{\beta^T \mathbf{Z}(k, l, \mathbf{x}_j)} P_{pB}(\boldsymbol{\ell}_k) \pi_k \eta_j.$$

Plugging η_j back into $\tilde{l}_r(\boldsymbol{\theta}, \boldsymbol{\eta})$, we see that the last two terms that involve η_j become

$$-n_{+j} \log \left\{ 1 + n_1 (n_0 \nu)^{-1} \sum_{k,l} e^{\beta^T \mathbf{Z}(k, l, \mathbf{x}_j)} P_{pB}(\boldsymbol{\ell}_k) \pi_k \right\} - n_1 \log \nu. \text{ Suppose that the conditional distribution of } (\mathbf{R}, Y) \text{ given } \mathbf{X} \text{ is characterized by}$$

$$\frac{\sum_{k,l} P_{\mu, \sum}(\mathbf{R} Y, k, l, \mathbf{X}) e^{\beta^* T \mathbf{Z}^*(k, l, \mathbf{X}, Y)} P_{pB}(\boldsymbol{\ell}_k) \pi_k}{\sum_{y=0,1} \sum_{k,l} e^{\beta^* T \mathbf{Z}^*(k, l, \mathbf{X}, y)} P_{pB}(\boldsymbol{\ell}_k) \pi_k},$$

where $\boldsymbol{\beta}^* = (\log\{(n_0 \nu)^{-1} n_1\}, \boldsymbol{\beta}^T)^T$, $\mathbf{Z}^*(k, l, \mathbf{x}, y) = (y, y \mathbf{Z}(k, l, \mathbf{x})^T)^T$. We can show that $\tilde{l}_r(\boldsymbol{\theta}, \boldsymbol{\eta})$ is equivalent to the log-likelihood

$$l_r^*(\boldsymbol{\theta}, \nu) = \sum_i \log \left\{ \frac{\sum_{k,l} P_{\mu, \sum}(\mathbf{R}_i Y_i, k, l, \mathbf{X}_i) e^{\beta^* T \mathbf{Z}^*(k, l, \mathbf{X}_i, Y_i)} P_{pB}(\boldsymbol{\ell}_k) \pi_k}{\sum_{y=0,1} \sum_{k,l} e^{\beta^* T \mathbf{Z}^*(k, l, \mathbf{X}_i, y)} P_{pB}(\boldsymbol{\ell}_k) \pi_k} \right\}.$$

We then maximize $l_r^*(\boldsymbol{\theta}, \nu)$ through the EM algorithm given below. The estimation of the covariance matrix of $(\boldsymbol{\theta}, \nu)$ is based on the information matrix of $l_r^*(\boldsymbol{\theta}, \nu)$.

The complete-data score function pertaining to $l_r^*(\boldsymbol{\theta}, \nu)$ is

$$\sum_i \log \left\{ \frac{\sum_{k,l} I(K_i=k, L_i=l) P_{\mu, \sum}(\mathbf{R}_i Y_i, k, l, \mathbf{X}_i) e^{\beta^* T Z^*(k,l, \mathbf{X}_i, Y_i)} P_{pB}(\ell_k) \pi_k}{\sum_{y=0,1} \sum_{k,l} e^{\beta^* T Z^*(k,l, \mathbf{X}_i, y)} P_{pB}(\ell_k) \pi_k} \right\}.$$

In the E-step, we calculate $P(K_i = k, L_i = l | \mathbf{R}_i, Y_i, \mathbf{X}_i)$, which can be shown to be

$$\omega_{ikl} \equiv \frac{P_{\mu, \sum}(\mathbf{R}_i Y_i, k, l, \mathbf{X}_i) e^{\beta^* T Z^*(k,l, \mathbf{X}_i, Y_i)} P_{pB}(\ell_k) \pi_k}{\sum_{k',l'} P_{\mu, \sum}(\mathbf{R}_i Y_i, k', l', \mathbf{X}_i) e^{\beta^* T Z^*(k',l', \mathbf{X}_i, Y_i)} P_{pB}(\ell_{k'}) \pi_{k'}}.$$

In the M-step, we use the Newton-Raphson algorithm. The rest of the algorithm is similar to that of Appendix A.2 except that for the starting values, we set $\beta^* = 0$, $\pi = (1/(1 + N_K), \dots, 1/(1 + N_K))$, and p_B to be the frequency of B allele in the annotation file.

In the absence of $\mathbf{X}$, we directly maximize (5) by an EM algorithm in which the complete-data score function is

$$\sum_i \log \left\{ \frac{\sum_{k,l} I(K_i=k, L_i=l) P_{\mu, \sum}(\mathbf{R}_i Y_i, k, l) e^{Y_i \beta^T Z(k,l)} P_{pB}(\ell_k) \pi_k}{\sum_{k,l} e^{Y_i \beta^T Z(k,l)} P_{pB}(\ell_k) \pi_k} \right\},$$

and the E-step is to evaluate

$$\omega_{ikl} \equiv P(K_i=k, L_i=l | \mathbf{R}_i, Y_i) = \frac{P_{\mu, \sum}(\mathbf{R}_i Y_i, k, l) e^{Y_i \beta^T Z(k,l)} P_{pB}(\ell_k) \pi_k}{\sum_{k',l'} P_{\mu, \sum}(\mathbf{R}_i Y_i, k', l') e^{Y_i \beta^T Z(k',l')} P_{pB}(\ell_{k'}) \pi_{k'}}.$$

The rest of the EM-algorithm is the same as above.

A.4 Pre-Processing of the Schizophrenia Data

We used the command “apt-probeset-summarize” of Affymetrix Power Tools to generate the allele-specific intensities from Affymetrix CEL files. We then followed the data-preprocessing protocol of PennCNV, called PennCNV-Affy, to convert the Affymetrix data to the Illumina format. To derive the starting values for (μ, Σ) , we used PennCNV to get CNV calls and then applied thresholds for the BAF measurements to obtain ASCN calls at SNP loci. If an ASCN state, e.g., $(K = 3, L = 1)$, has less than five individuals, we excluded that state and the corresponding individuals from the analysis as the value of (μ, Σ) for that state can not be estimated robustly.

Appendix B: Asymptotic Properties

B.1 Sieve MLE of Likelihood (4)

We first state the identifiability of θ and G in Lemma 1.

LEMMA 1. If two sets of parameters (θ, G) and $(\tilde{\theta}, \tilde{G})$ yield the same likelihood, then $\theta = \tilde{\theta}$ and $G = \tilde{G}$.

Proof: Suppose that

$$\sum_{k,l} P_{\mu, \Sigma}(\mathbf{R}|Y, k, l, \mathbf{X}) P_{\alpha, \beta}(Y|k, l, \mathbf{X}) G(k, l|\mathbf{X}) = \sum_{k,l} P_{\tilde{\mu}, \tilde{\Sigma}}(\mathbf{R}|Y, k, l, \mathbf{X}) P_{\tilde{\alpha}, \tilde{\beta}}(Y|k, l, \mathbf{X}) \tilde{G}(k, l|\mathbf{X}).$$

By Proposition 1 of Teicher (1963), which states that all finite mixtures of normal distributions are identifiable, we see that $\mu = \tilde{\mu}$ and $\Sigma = \tilde{\Sigma}$ and that for all (k, l) , $P_{\alpha, \beta}(Y|k, l, \mathbf{X}) G(k, l|\mathbf{X}) = P_{\tilde{\alpha}, \tilde{\beta}}(Y|k, l, \mathbf{X}) \tilde{G}(k, l|\mathbf{X})$. The summation over $Y = 0$ and 1 yields $G = \tilde{G}$. Thus, $P_{\alpha, \beta}(Y|k, l, \mathbf{X}) = P_{\tilde{\alpha}, \tilde{\beta}}(Y|k, l, \mathbf{X})$. Because $P_{\alpha, \beta}(Y|k, l, \mathbf{X})$ is a logistic regression model with a full-rank design matrix, $\alpha = \tilde{\alpha}$ and $\beta = \tilde{\beta}$.

We state in Lemma 2 that the information matrices along all non-trivial parametric submodels are non-singular.

LEMMA 2. If there exist a vector $\mathbf{v}_\theta = (v_\alpha \ v_\beta \ v_\mu \ v_\Sigma)$ and a function $\psi(k, l, \mathbf{x})$ with $E[\psi(K, L, \mathbf{X})|\mathbf{X}] = 0$ such that

$$\mathbf{v}_\theta^T l_\theta(\theta_0, G_0) + l_G(\theta_0, G_0)[\psi] = 0, \quad (9)$$

where l_θ is the score function for θ , and $l_G[\psi]$ is the score function for G along the submodel $G_0(1 + \epsilon\psi)$, then $\mathbf{v}_\theta = \mathbf{0}$ and $\psi = 0$.

Proof: We focus on likelihood (6), which is the univariate version of likelihood (4). Equation (9) can be expanded as

$$\begin{aligned} \sum_{k=0}^{N_K} P_{\mu, \Sigma}(\mathbf{R}|Y, k, \mathbf{X}) \left(P_{\alpha, \beta}(Y|k, \mathbf{X}) G(k|\mathbf{X}) \left[\frac{(R - \mu_{Y, k, \mathbf{X}})}{\sigma_{Y, k, \mathbf{X}}^2} \nu_{\mu_{Y, k, \mathbf{X}}} - \frac{1}{2} \left\{ 1 - \frac{(R - \mu_{Y, k, \mathbf{X}})^2}{\sigma_{Y, k, \mathbf{X}}^2} \right\} \nu_{\sigma_{Y, k, \mathbf{X}}^2} \right] \right. \\ \left. + \mathbf{v}_{\alpha, \beta}^T \nabla_{\alpha, \beta} P_{\alpha, \beta}(Y|k, \mathbf{X}) G(k|\mathbf{X}) + P_{\alpha, \beta}(Y|k, \mathbf{X}) G(k|\mathbf{X}) \psi(k, \mathbf{X}) \right) = 0. \end{aligned} \quad (10)$$

Here and in the sequel, $\nabla_{\mathbf{u}} f(\mathbf{u}, \mathbf{v}) = f(\mathbf{u}, \mathbf{v}) / \mathbf{u}$. Equation (10) is essentially

$$\sum_{k=0}^{N_K} \exp \left\{ -\frac{(R - \mu_k)^2}{2\sigma_k^2} \right\} (a_k R^2 + b_k R + c_k) = 0, \quad (11)$$

where $\mu_k = \mu_{Y,k,\mathbf{X}}$ and $\sigma_k^2 = \sigma_{Y,k,\mathbf{X}}^2$. Write $\phi_k = \exp \left\{ -(R - \mu_k)^2 / (2\sigma_k^2) \right\}$ and reorder the component $\phi_0, \dots, \phi_{N_K}$ lexicographically by: $\phi_k \prec \phi_{k'}$ if $\sigma_k > \sigma_{k'}$ or if $\sigma_k = \sigma_{k'}$ but $\mu_k > \mu_{k'}$. We denote the sequence of subscripts of the ordered ϕ_k 's by $(0, \dots, \tilde{N}_K)$. Dividing (11) by ϕ_0 , we have

$$a_0 R^2 + b_0 R + c_0 = - \sum_{k=1}^{\tilde{N}_K} \exp \left\{ -\frac{(R - \mu_k)^2}{2\sigma_k^2} + \frac{(R - \mu_0)^2}{2\sigma_0^2} \right\} (a_k R^2 + b_k R + c_k).$$

Each exponential component on the right-hand side is of the order $\exp(-R^2)$ or $\exp(-R)$, so the right-hand side goes to zero as $R \rightarrow \infty$. By contrast, the left-hand side goes to infinity unless $a_0 = 0$ and $b_0 = 0$ are both 0. Thus $a_0 = 0$, $b_0 = 0$ and $c_0 = 0$. Divide the remainder of (11) by ϕ_1 . By the same argument, $a_1 = 0$, $b_1 = 0$ and $c_1 = 0$. Thus, a_k , b_k and c_k are 0 for all k , which implies that $\nu_{\mu, \Sigma} = 0$ and

$\nu_{\alpha, \beta}^T \nabla_{\alpha, \beta} P_{\alpha, \beta}(Y|k, \mathbf{X}) G(k|\mathbf{X}) + P_{\alpha, \beta}(Y|k, \mathbf{X}) G(k|\mathbf{X}) \psi(k, \mathbf{X}) = 0$ for all k . Assuming that the function $G(k|\mathbf{x})$ is positive in its support for all $\mathbf{x}$, we sum over $Y = 0$ and 1 to yield $\psi = 0$.

Thus, $\nu_{\alpha, \beta}^T \nabla_{\alpha, \beta} P_{\alpha, \beta}(Y|k, \mathbf{X}) = 0$ for all k . It then follows that $\nu_{\alpha, \beta} = 0$ under the logistic regression model $P_{\alpha, \beta}(Y|k, \mathbf{X})$ with a full-rank design matrix.

We now prove Theorem 1. To simplify notation, we represent (K, L) by an ASCN state index S taking values $1, \dots, N_S$, where N_S is the total number of states. Write $\mathbf{O} = (\mathbf{R}, Y)$ and let $P_{\theta}(\mathbf{O}|\mathbf{X}, S)$ be the conditional density of $\mathbf{O}$ given $(\mathbf{X}, S)$, where $\theta = (\alpha, \beta, \mu, \Sigma)$.

Proof of Theorem 1: We first consider the case of $d \equiv \dim(\mathbf{X}) = 1$. The case of $d > 1$ is proven at the end. In the sequel, we denote $\mathbf{P}f = \int f d\mathbf{P}$, $\mathbf{P}_n f = n^{-1} \sum_{i=1}^n f(\mathbf{O}_i, X_i)$ and $\mathbf{G}_n f = \sqrt{n}(\mathbf{P}_n f - \mathbf{P}f)$.

We first prove the consistency of $(\hat{\theta}, \hat{G})$. Let $l(\theta, G) = \log \left\{ \sum_{s=1}^{N_S} P_{\theta}(\mathbf{O}|X, s) G(s|X) \right\}$. The

sieve space for $G(s|x)$ is $S_n = \left\{ G(s|x) = \sum_{m=1}^{M_n} g_m(s) I(x \in \mathcal{X}_m) \right\}$, where the $g_m(s)$'s are probability mass functions, and $(\mathcal{X}_1, \dots, \mathcal{X}_{M_n})$ is an equal partition of X 's support and the diameters are bounded by $C M_n^{-1/d}$ for some constant C . By definition, the sieve estimator $(\hat{\theta}, \hat{G})$ maximizes $\mathbf{P}_n l(\theta, G)$ over $\theta \in \Theta$ and $G \in S_n$. We denote the true conditional probability function $G(S = s|\mathbf{X} = \mathbf{x})$ by $G_0(s|\mathbf{x})$, and define

$\tilde{G}(s|x) = \sum_{m=1}^{M_n} I(x \in \mathcal{X}_m) |\mathcal{X}_m|^{-1} \int_{\mathcal{X}_m} G_0(s|x) dx$, where $|\mathcal{X}_m|$ is the volume of $\mathcal{X}_m$. We assume that $G_0(s|\mathbf{x})$ is continuously differentiable and uniformly bounded away from zero and that the support of $\mathbf{X}$ is bounded. Then $\sup_x |G(s|x) - \tilde{G}(s|x)| \leq C_1 M_n^{-1/d}$ for some constant C_1 , implying that G is bounded away from zero uniformly in x . We further define $G^*(s|x) = \{ \tilde{G}(s|x) + G(s|x) \} / 2$, so that $\tilde{G}^*(s|x)$ is also bounded away from zero. Since $\mathbf{P}_n l(\hat{\theta}, \hat{G}) \rightarrow \mathbf{P}_n l(\hat{\theta}_0, G)$, the concavity of $l(\theta, G)$ in G yields $\mathbf{P}_n l(\hat{\theta}, \hat{G}^*) \rightarrow \mathbf{P}_n l(\hat{\theta}_0, G)$, i.e.,

$$\mathbf{G}_n \left\{ l(\hat{\boldsymbol{\theta}}, \hat{G}^*) - l(\boldsymbol{\theta}_0, \tilde{G}) \right\} \geq \sqrt{n} \mathbf{P} l(\boldsymbol{\theta}_0, \tilde{G}) - \sqrt{n} \mathbf{P} l(\hat{\boldsymbol{\theta}}, \hat{G}^*). \quad (12)$$

We examine the left-hand side of (12). Note that $l(\boldsymbol{\theta}, G)$ is a random variable indexed by $(\boldsymbol{\theta}, g_m(s))$ and is Lipschitz continuous with respect to $(\boldsymbol{\theta}, g_m(s))$. Because $P_{\boldsymbol{\theta}}(\mathbf{O}|\mathbf{X}, s)$ is the product of a (truncated) normal density function and a logistic function, the Lipschitz coefficient is bounded by some function $H(\mathbf{O}, \mathbf{X})$ with a finite second moment. Thus, the ϵ -bracket number for $l(\boldsymbol{\theta}, G)$ is of order $O(\epsilon^{-q-M_n N_S})$, where $q = \dim(\boldsymbol{\theta})$. In addition, $l(\hat{\boldsymbol{\theta}}, \hat{G}^*) - l(\boldsymbol{\theta}_0, \tilde{G})$ has a bounded envelope function. By the large deviation inequality for empirical processes (van der Vaart 1996, Thm 19.20.1), the left-hand side of (12) is

$$\mathbf{G}_n \left\{ l(\hat{\boldsymbol{\theta}}, \hat{G}^*) - l(\boldsymbol{\theta}_0, \tilde{G}) \right\} = \int_0^{O(1)} \sqrt{\log \epsilon^{-q-M_n N_S}} d\epsilon = O(\sqrt{M_n}). \quad (13)$$

We examine the right-hand side of (12). Since G_0 maximizes $\mathbf{P} l(\boldsymbol{\theta}_0, G)$, we have

$\mathbf{P} l(\boldsymbol{\theta}_0, G_0) + O(\|\tilde{G} - G_0\|_{L_2(P)}^2) = \mathbf{P} l(\boldsymbol{\theta}_0, G_0) + O(M_n^{-2/d})$, which, together with (12) and (13), yields $O(\sqrt{M_n}/\sqrt{n}) + O(M_n^{-2/d}) \geq \mathbf{P} l(\boldsymbol{\theta}_0, G_0) - \mathbf{P} l(\hat{\boldsymbol{\theta}}, \hat{G}^*)$. Thus, the Kulback-Leibler (KL) distance between $(\boldsymbol{\theta}_0, G_0)$ and $(\hat{\boldsymbol{\theta}}, \hat{G}^*)$ converges to zero. By the relationship between the KL and Hellinger distance and the fact that the density indexed by these parameters is positive, we have

$E \left[\left(\sum_{s=1}^{N_S} P_{\hat{\boldsymbol{\theta}}}(\mathbf{O}|X, s) \hat{G}^*(\mathbf{O}|X) - \sum_{s=1}^{N_S} P_{\boldsymbol{\theta}_0}(\mathbf{O}|X, s) G_0(s|X) \right)^2 \right] \rightarrow_p 0$, which also holds when $\hat{G}^*$ is replaced by $\hat{G}$. Thus, for almost every $(\mathbf{o}, x)$,

$\sum_{s=1}^{N_S} P_{\hat{\boldsymbol{\theta}}}(\mathbf{o}|x, s) \hat{G}(\mathbf{o}|x) - \sum_{s=1}^{N_S} P_{\boldsymbol{\theta}_0}(\mathbf{o}|x, s) G_0(\mathbf{o}|x) \rightarrow 0$. By the subsequence arguments and the identifiability result in Lemma 1, we obtain $\hat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}_0$ and $\hat{G}(s|x) \rightarrow G_0(s|x)$. We expand $\mathbf{P} l(\hat{\boldsymbol{\theta}}, \hat{G}^*)$ in the neighborhood of $(\boldsymbol{\theta}_0, G_0)$ as

$\mathbf{P} l(\boldsymbol{\theta}_0, G_0) + \frac{\partial^2}{\partial t^2} l(\boldsymbol{\theta}_0 + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), G_0 + t(\hat{G}^* - G_0)) \big|_{t=t^*}$, where $t^* \in (0, 1)$. For our model of $l(\boldsymbol{\theta}, G)$, there exists some positive constant C_2 such that the foregoing second derivative is bounded by $-C_2 \left\{ \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2 + \|\hat{G}^* - G_0\|_{L_2(P)}^2 \right\}$. This means

$$\sqrt{n} \mathbf{P} l(\boldsymbol{\theta}_0, \tilde{G}) - \sqrt{n} \mathbf{P} l(\hat{\boldsymbol{\theta}}, \hat{G}^*) \geq C_2 \sqrt{n} \left\{ \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2 + \|\hat{G} - G_0\|_{L_2(P)}^2 \right\} + O(\sqrt{n} M_n^{-2/d}). \quad (14)$$

Combining (13) and (14), we obtain

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2 + \|\hat{G} - G_0\|_{L_2(P)}^2 = O_p(\sqrt{M_n}/\sqrt{n}) + O(M_n^{-2/d}), \quad (15)$$

from which the consistency is established.

The convergence rate in (15) is not sufficient for the derivation of asymptotic normality. We improve the convergence rate by revisiting (12). By (15), the envelope function for $\{l(\hat{\boldsymbol{\theta}}, \hat{G}^*)\}$

$-l(\theta_0, G)$ has the second moment bounded by $c_n \equiv M_n^{1/4}/n^{1/4} + O(M_n^{-1/d})$. Thus, the use of the large deviation inequality for empirical processes implies that the left-hand side of (12) is bounded by $O\left(\sqrt{M_n}\right) c_n \log c_n = o(1)$. By applying previous arguments to the right-hand side of (12), we obtain an improved convergence rate $o_p(1/\sqrt{n}) + O(M_n^{-2/d})$ for equation (15).

To prove the asymptotic efficiency, we first determine the least favorable direction for G . That is, we determine $\psi_0(s, x)$ such that $E\{\psi_0(S, X)|X\} = 0$ and

$$E\left[\{l_\theta(\theta_0, G_0) - l_G(\theta_0, G_0)[\psi_0]\} l_G[\tilde{\psi}]\right] = 0 \quad (16)$$

for any $\tilde{\psi}(s, x)$, where $l_\theta(\theta_0, G_0) \equiv E\{\log P_\theta(\mathbf{O}|X, S)/\theta(\mathbf{O}, X)\}$ is the score function for θ and $l_G(\theta_0, G_0)[\psi] \equiv E[\psi(S, X)|\mathbf{O}, X]$ is the score function along the submodel $G_0(s|x)\{1 + \epsilon\psi(s, x)\}$. Equation (16) is equivalent to

$$E\left[E\{\partial \log P_\theta(\mathbf{O}|X, S)/\partial \theta - \psi_0(S, X)|\mathbf{O}, X\} \tilde{\psi}(S, X)\right] = 0, \text{ which implies that}$$

$E[E\{\partial \log P_\theta(\mathbf{O}|X, S)/\partial \theta - \psi_0(S, X)|\mathbf{O}, X\}|S, X] = a(X)$ for some function $a(\cdot)$. Since S is discrete, this yields a linear equation system

$$\begin{aligned} \sum_{s=1}^{N_S} E[P(S=s|\mathbf{O}, X) E\{\partial \log P_\theta(\mathbf{O}|X, S)/\partial \theta | \mathbf{O}, X\} | S=t, X] P(S=t|X) &= \sum_{s=1}^{N_S} \psi_0(s, X) E\{P(S \\ &= s|\mathbf{O}, X) P(S \\ &= t|\mathbf{O}, X)|X\} + a(X) P(S=t|X). \end{aligned}$$

It is easy to show that the matrix $(E\{P(S=s|\mathbf{O}, X) P(S=t|\mathbf{O}, X)|X\})_{s,t}$ is non-singular. Suppose that there exists a vector $(a_1, \dots, a_{N_S})$ such that

$$(a_1, \dots, a_{N_S}) (E\{P(S=s|\mathbf{O}, X) P(S=t|\mathbf{O}, X)|X\})_{s,t} (a_1, \dots, a_{N_S})^T = 0,$$

which is equivalent to

$$E\left[\left\{\sum_{s=1}^{N_S} a_s P(S=s|\mathbf{O}, X)\right\}^2\right] = 0.$$

Thus, $\sum_{s=1}^{N_S} a_s P(S=s|\mathbf{O}, X) = 0$ almost surely. This implies that

$\sum_{s=1}^{N_S} a_s P(\mathbf{O}|S=s, X) P(S=s|X) = 0$. Following the proof of Lemma 1, $a_1 = \dots = a_{N_S} = 0$. The non-singularity means that there exists a unique solution $\psi_0(s, x)$ to $E\{\psi_0(S, X)|X\} = 0$. Clearly, $\psi_0(s, x)$ is continuously differentiable with respect to x . We define

$\tilde{\psi}_0(s, x) = \sum_{m=1}^{M_n} I(x \in \mathcal{X}_m) |\mathcal{X}_m|^{-1} \int_{\mathcal{X}_m} \psi_0(s, x) dx$, so that $\tilde{\psi}_0(s, x)$ is on the tangent space of S_n . Since $(\hat{\theta}, \hat{G})$ maximizes $\mathbf{P}_n l(\theta, G)$, we have $\mathbf{P}_n l'(\hat{\theta}, \hat{G}) = 0$ and $\mathbf{P}_n l'_G(\hat{\theta}, \hat{G})[\tilde{\psi}_0] = 0$. Thus,

$$\mathbf{G}_n = \{l_{\theta}(\hat{\theta}, \hat{G}) - l_G(\hat{\theta}, \hat{G})[\tilde{\psi}_0]\} = -\sqrt{n} \mathbf{P} \{l_{\theta}(\hat{\theta}, \hat{G}) - l_G(\hat{\theta}, \hat{G})[\tilde{\psi}_0]\}. \quad (17)$$

By the property of $\tilde{\psi}_0$, we expand the right-hand side of (17) as

$$-\sqrt{n} \mathbf{P} (l_{\theta\theta} - l_{G\theta}[\psi_0]) (\hat{\theta} - \theta_0) + \sqrt{n} O \left(\|\hat{\theta} - \theta_0\|^2 + \|\hat{G} - G_0\|_{L_2(P)}^2 + \|\psi_0 - \tilde{\psi}_0\|_{L_2(P)}^2 \right),$$

the second term of which becomes $o_p(1) + O(\sqrt{n} M_n^{-2/d}) = o_p(1)$ by the improved convergence rate and the choice of number of nodes $M_n = n^\tau$, where $\tau \in (1/4, 1/3)$. For the left-hand side of (17), we use the similar arguments in deriving the improved rate to get

$$\mathbf{G}_n \left\{ (l_{\theta}(\hat{\theta}, \hat{G}) - l_G(\hat{\theta}, \hat{G})[\tilde{\psi}_0]) - (l_{\theta}(\theta_0, G_0) - l_G(\theta_0, G_0)[\psi_0]) \right\} = o_p(1). \text{ Combining the results of both sides of (17), we have}$$

$$\mathbf{G}_n \{l_{\theta}(\theta_0, G_0) - l_G(\theta_0, G_0)[\psi_0]\} = -\sqrt{n} \mathbf{P} (l_{\theta\theta} - l_{G\theta}[\psi_0]) (\hat{\theta} - \theta_0) + o_p(1).$$

Lemma 2 implies the non-singularity of $\mathbf{P}(l_{\theta\theta} - l_{G\theta}[\psi_0])$, so $\hat{\theta}$ is an asymptotically linear estimator of θ_0 whose influence function is exactly the efficient influence function. Hence, $\sqrt{n}(\hat{\theta} - \theta_0)$ converges in distribution to a mean-zero normal vector whose covariance matrix attains the semiparametric efficiency bound.

When $d > 1$, the histogram sieve no longer works because the bias due to the histogram can only be of order $M_n^{-1/d}$, which may dominate the variability. In this case, we consider the sieve estimation based on general B-splines described in Section 2.2. Now we can control the bias of order $M_n^{-r/d}$, where r is the smoothness of $G_0(s|\mathbf{x})$, which is assumed to belong to the Sobolev space $W^{r,\infty}$ with $r > d/2$. The arguments for $d = 1$ yields

$$\mathbf{G}_n (l_{\theta}(\theta_0, G_0) - l_G(\theta_0, G_0)[\psi_0]) = -\sqrt{n} \mathbf{P} (l_{\theta\theta} - l_{G\theta}[\psi_0]) (\hat{\theta} - \theta_0) + o_p(1) + O_p(\sqrt{n} M_n^{-2r/d}).$$

Note that d is replaced by d/r . We set $M_n = n^\tau$, where $\tau \in (d/4r, 1/3)$. Then

$$\sqrt{n} M_n^{-2r/d} \rightarrow 0, \text{ so the asymptotic properties of Theorem 1 continue to hold.}$$

B.2 NPMLE of Likelihood (5)

We first state the identifiability of θ and F in Lemma 3.

LEMMA 3. If two sets of parameters (θ, F) and $(\tilde{\theta}, \tilde{F})$ yield the same likelihood, then $\theta = \tilde{\theta}$ and $F = \tilde{F}$.

Proof: Suppose that

$$\sum_{k,l} P_{\mu, \Sigma}(\mathbf{R}^Y, k, l, \mathbf{X}) \frac{e^{Y\beta^T Z(k,l,\mathbf{X})} P_{p_B}(\mathbf{l}_k) \pi_k f(\mathbf{X})}{\sum_{k',l'} \int_{\mathbf{x}} e^{Y\beta^T Z(k',l',\mathbf{x})} P_{p_B}(\mathbf{l}_{k'}) \pi_{k'} dF(\mathbf{x})} = \sum_{k,l} P_{\tilde{\mu}, \tilde{\Sigma}}(\mathbf{R}^Y, k, l, \mathbf{X}) \frac{e^{Y\tilde{\beta}^T Z(k,l,\mathbf{X})} P_{\tilde{p}_B}(\mathbf{l}_k) \tilde{\pi}_k \tilde{f}(\mathbf{X})}{\sum_{k',l'} \int_{\mathbf{x}} e^{Y\tilde{\beta}^T Z(k',l',\mathbf{x})} P_{\tilde{p}_B}(\mathbf{l}_{k'}) \tilde{\pi}_{k'} d\tilde{F}(\mathbf{x})}. \quad (18)$$

Letting $Y = 0$ yields $\sum_{k,l} P_{\mu, \Sigma}(\mathbf{R}|Y, k, l, \mathbf{X}) P_{p_B}(\mathbf{l}_k) \pi_k f(\mathbf{X}) = \sum_{k,l} P_{\tilde{\mu}, \tilde{\Sigma}}(\mathbf{R}|Y, k, l, \mathbf{X}) P_{\tilde{p}_B}(\mathbf{l}_k) \tilde{\pi}_k \tilde{f}(\mathbf{X})$. By Proposition 1 of Teicher (1963), we have $\mu = \tilde{\mu}$, $\Sigma = \tilde{\Sigma}$, and $P_{p_B}(\mathbf{l}_k) \pi_k f(\mathbf{X}) = P_{\tilde{p}_B}(\mathbf{l}_k) \tilde{\pi}_k \tilde{f}(\mathbf{X})$ for all (k, l) . The last equation implies that $f = \tilde{f}$, $\pi_k = \tilde{\pi}_k$ and $p_B = \tilde{p}_B$. Letting $Y = 1$ in (18) and applying Proposition 1 of Teicher (1963) again, we see that

$$e^{\beta^T Z(k,l,\mathbf{X})} \left\{ \sum_{k',l'} \int_{\mathbf{x}} e^{\beta^T Z(k',l',\mathbf{x})} P_{p_B}(\mathbf{l}_{k'}) \pi_{k'} dF(\mathbf{x}) \right\}^{-1} = e^{\tilde{\beta}^T Z(k,l,\mathbf{X})} \left\{ \sum_{k',l'} \int_{\mathbf{x}} e^{\tilde{\beta}^T Z(k',l',\mathbf{x})} P_{\tilde{p}_B}(\mathbf{l}_{k'}) \tilde{\pi}_{k'} d\tilde{F}(\mathbf{x}) \right\}^{-1}$$

for all (k, l) . It then follows from the linear independence of $\{1, Z(k, l, \mathbf{X})^T\}$ that $\beta = \tilde{\beta}$.

Next, we state in Lemma 4 that the information matrices along all non-trivial parametric submodels are non-singular.

LEMMA 4. If there exist a vector $\nu_{\theta} = (\nu_{\beta} \ \nu_{\pi} \ \nu_{p_B} \ \nu_{\mu} \ \nu_{\Sigma})$ and a function $\psi(\mathbf{x})$ with $E[\psi(\mathbf{X})] = 0$ such that

$$\nu_{\theta}^T l_{\theta}(\theta_0, F_0) + l_F(\theta_0, F_0) \left[\int \psi dF_0 \right] = 0, \quad (19)$$

where l_{θ} is the score function for θ , and $l_F[\int \psi dF_0]$ is the score function for F along the submodel $F_0 + \epsilon \int \psi dF_0$, then $\nu_{\theta} = 0$ and $\psi = 0$.

Proof: We first set $Y = 0$ in (19), which becomes

$$\sum_{k,l} P_{\mu, \Sigma}(\mathbf{R}^Y, k, l, \mathbf{X}) \left\{ \nu_{\mu, \Sigma}^T \nabla_{\mu, \Sigma} \log P_{\mu, \Sigma}(\mathbf{R}^Y, k, l, \mathbf{X}) P_{p_B}(\mathbf{l}_k) \pi_k + \nu_{\pi, p_B}^T \nabla_{\pi, p_B} P_{p_B}(\mathbf{l}_k) \pi_k + P_{p_B}(\mathbf{l}_k) \pi_k \psi(\mathbf{X}) \right\} = 0.$$

By the arguments in the proof of Lemma 2, $\nu_{\mu, \Sigma} = 0$ and for all (k, l) ,

$$\nu_{\pi, p_B}^T \nabla_{\pi, p_B} P_{p_B}(\mathbf{l}_k) \pi_k + P_{p_B}(\mathbf{l}_k) \pi_k \psi(\mathbf{X}) = 0. \text{ Summing over } (k, l) \text{ yields } \psi = 0 \text{ and then } \nu_{\pi, p_B} = 0. \text{ Letting } Y = 1 \text{ in (19), we have}$$

$$\begin{aligned} & \sum_{k,l} P_{\mu, \Sigma}(\mathbf{R}^Y, k, l, \mathbf{X}) e^{\beta^T Z(k,l,\mathbf{X})} P_{p_B}(\mathbf{l}_k) \pi_k \\ & \times \left\{ \nu_{\beta}^T Z(k, l, \mathbf{X}) - \frac{\sum_{k'} \sum_{l'} \int_{\mathbf{x}} e^{\beta^T Z(k',l',\mathbf{x})} \nu_{\beta}^T Z(k', l', \mathbf{x}) P_{p_B}(\mathbf{l}_{k'}) \pi_{k'} dF(\mathbf{x})}{\sum_{k'} \sum_{l'} \int_{\mathbf{x}} e^{\beta^T Z(k',l',\mathbf{x})} P_{p_B}(\mathbf{l}_{k'}) \pi_{k'} dF(\mathbf{x})} \right\} = 0. \end{aligned}$$

Proposition 1 of Teicher (1963) implies that $\nu_{\beta}^T Z(k, l, \mathbf{X})$ is a constant for all (k, l) . Because the vector $\{1, Z(K, L, \mathbf{X})^T\}$ is linearly independent, $\nu_{\beta} = 0$.

Finally, we prove Theorem 2.

Proof of Theorem 2: The likelihood given in (5) resembles (2.5) and (2.8) of Hu et al. (2010), so the arguments in the proofs of their Theorem S.1 and S.2 can be used to prove Theorem 2 with some modifications. Specifically, the score equation that the Lagrange multiplier $\hat{\lambda}$ satisfies becomes

$$\frac{1}{\hat{F}\{\mathbf{X}_i\}} - \frac{I(Y_i=1) \sum_{k,l} e^{\beta^T Z(k,l,\mathbf{X}_i)} P_{\hat{p}_B}(\mathbf{k}) \hat{\pi}_k}{\sum_{k,l} \int_{\mathbf{x}} e^{\beta^T Z(k,l,\mathbf{x})} P_{\hat{p}_B}(\mathbf{k}) \hat{\pi}_k d\hat{F}(\mathbf{x})} - \hat{\lambda} = 0,$$

where $F\{\mathbf{X}_i\}$ is the jump size of F at $\mathbf{X}_i$, and consequently, $\hat{\lambda} \doteq n - n_1$. The definition of F in Hu et al. (2010) is replaced by

$$\tilde{F}\{\mathbf{X}_i\} = \left\{ \frac{I(Y_i=1) \sum_{k,l} e^{\beta_0^T Z(k,l,\mathbf{X}_i)} P_{p_{B,0}}(\mathbf{k}) \pi_{k,0}}{\sum_{k,l} \int_{\mathbf{x}} e^{\beta_0^T Z(k,l,\mathbf{x})} P_{p_{B,0}}(\mathbf{k}) \pi_{k,0} dF_0(\mathbf{x})} + n - n_1 \right\}^{-1}.$$

Then, the consistency follows from the identifiability given in Lemma 3 and the arguments in the proof of Theorem S.1 of Hu et al. (2010). The asymptotic normality and efficiency follow from Lemma 4 and the arguments in the proof of Theorem S.2 of Hu et al. (2010).

References

- Amos CI, Wu XF, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, et al. Genome-Wide Association Scan of Tag SNPs Identifies A Susceptibility Locus for Lung Cancer at 15q25.1. *Nature Genetics*. 2008; 40:616–622. [PubMed: 18385676]
- Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, Hurles ME. A Robust Statistical Method for Case-Control Association Testing With Copy Number Variation. *Nature Genetics*. 2008; 40:1245–1252. [PubMed: 18776912]
- Bucan M, Abrahams BS, Wang K, Glessner JT, Herman EI, et al. Genome-Wide Analyses of Exonic Copy Number Variants in A Family-Based Study Point to Novel Autism Susceptibility Genes. *PLoS Genetics*. 2009; 5:e1000536. [PubMed: 19557195]
- Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J. QuantiSNP: An Objective Bayes Hidden-Markov Model to Detect and Accurately Map Copy Number Variation Using SNP Genotyping Data. *Nucleic Acids Research*. 2007; 35:2013–2025. [PubMed: 17341461]
- Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, et al. Copy Number Variation at 1q21.1 Associated with Neuroblastoma. *Nature*. 2009; 459:987–991. [PubMed: 19536264]
- Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, et al. Autism Genome-Wide Copy Number Variation Reveals Ubiquitin and Neuronal Genes. *Nature*. 2009; 459:569–573. [PubMed: 19404257]
- Handsaker RE, Korn JM, Nemesh J, McCarroll SA. Discovery and Genotyping of Genome Structural Polymorphism by Sequencing on A Population Scale. *Nature Genetics*. 2011; 43:269–276. [PubMed: 21317889]

- Hu YJ, Lin DY. Analysis of Untyped SNPs: Maximum Likelihood and Imputation Methods. *Genetic Epidemiology*. 2010; 34:803–815. [PubMed: 21104886]
- Hu YJ, Lin DY, Zeng D. A General Framework for Studying Genetic Effects and Gene-Environment Interactions With Missing Data. *Biostatistics*. 2010; 11:583–598. [PubMed: 20348396]
- International Schizophrenia Consortium. Rare Chromosomal Deletions and Duplications Increase Risk of Schizophrenia. *Nature*. 2008; 455:237–241. [PubMed: 18668038]
- International Schizophrenia Consortium. Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder. *Nature*. 2009; 460:748–752. [PubMed: 19571811]
- Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, et al. Integrated Genotype Calling and Association Analysis of SNPs, Common Copy Number Polymorphisms and Rare CNVs. *Nature Genetics*. 2008; 40:1253–1260. [PubMed: 18776909]
- LaFramboise T, Weir BA, Zhao X, Beroukhir R, Li C, Harrington D, Sellers WR, Meyerson M. Allele-Specific Amplification in Cancer Revealed by SNP Array Analysis. *PLoS Computational Biology*. 2005; 1:e65. [PubMed: 16322765]
- Louis TA. Finding the Observed Information Matrix When Using the EM Algorithm. *Journal of the Royal Statistical Society, Series B*. 1982; 44:226–233.
- Lupski JR. Genomic Disorders Ten Years On. *Genome Medicine*. 2009; 1:42. [PubMed: 19439022]
- McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, et al. Integrated Detection and Population-Genetic Analysis of SNPs and Copy Number Variation. *Nature Genetics*. 2008; 40:1166–1174. [PubMed: 18776908]
- Medvedev P, Stanciu M, Brudno M. Computational Methods for Discovering Structural Variation With Next-Generation Sequencing. *Nature Methods*. 2009; 6:S13–S20. [PubMed: 19844226]
- Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping Copy Number Variation by Population-Scale Genome Sequencing. *Nature*. 2011; 470:59–65. [PubMed: 21293372]
- Murphy SA, van der Vaart AW. On Profile Likelihood. *Journal of the American Statistical Association*. 2000; 95:449–465.
- Need AC, Ge D, Weale ME, et al. A Genome-Wide Investigation of SNPs and CNVs in Schizophrenia. *PLoS Genetics*. 2009; 5:e1000373. [PubMed: 19197363]
- Prentice RL, Pyke R. Logistic Disease Incidence Models and Case- Control Studies. *Biometrika*. 1979; 66:403–441.
- Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global Variation in Copy Number in the Human Genome. *Nature*. 2006; 444:444–454. [PubMed: 17122850]
- Roeder K, Carroll RJ, Lindsay BG. A Semiparametric Mixture Approach to Case-Control Studies With Errors in Covariables. *Journal of the American Statistical Association*. 1996; 91:722–732.
- Shen X. On Methods of Sieves and Penalization. *The Annals of Statistics*. 1997; 25:2555–2591.
- Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, Dudbridge F, et al. Common Variants on Chromosome 6p22.1 Are Associated With Schizophrenia. *Nature*. 2009; 460:753–757. [PubMed: 19571809]
- Stefansson H, Ophoff RA, Steinberg S, et al. Common Variants Conferring Risk of Schizophrenia. *Nature*. 2009; 460:744–747. [PubMed: 19571808]
- Stefansson H, Rujescu D, Cichon S, et al. Large Recurrent Microdeletions Associated with Schizophrenia. *Nature*. 2008; 455:232–236. [PubMed: 18668039]
- Sun W, Wright FA, Tang Z, Nordgard SH, Van Loo P, Yu T, Kristensen VN, Perou CM. Integrated Study of Copy Number States and Genotype Calls Using High-Density SNP Arrays. *Nucleic Acids Research*. 2009; 37:5365–5377. [PubMed: 19581427]
- Teicher H. Identifiability of Finite Mixtures. *The Annals of Mathematical Statistics*. 1963; 34:1265–1269.
- Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE. Fine-Scale Structural Variation of the Human Genome. *Nature Genetics*. 2005; 37:727–732. [PubMed: 15895083]

- Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, et al. Allele-Specific Copy Number Analysis of Tumors. *Proceedings of the National Academy of Sciences*. 2010; 107:16910–16915.
- Wang K, Li M, Hadley D, Liu R, Glessner J, Grant S, Hakonarson H, Bucan M. PennCNV: An Integrated Hidden Markov Model Designed for High-Resolution Copy Number Variation Detection in Whole-Genome SNP Genotyping Data. *Genome Research*. 2007; 17:1665–1674. [PubMed: 17921354]
- Wang K, Zhang H, Ma D, Bucan M, Glessner JT, et al. Common Genetic Variants on 5p14.1 Associate with Autism Spectrum Disorders. *Nature*. 2009; 459:528–533. [PubMed: 19404256]

Figure 1.

Observed intensity measurements of 2550 individuals at SNP rs983985 in a schizophrenia study (Shi et al. 2009): (a) scatter plot of the Affymetrix data; (b) scatter plot of the LRR and BAF measurements derived from the Affymetrix data shown in (a) by the method of Wang et al. (2007); (c) density plot of the LRR measurements only. The color circles attempt to cluster the measurements into discrete ASCN/CNV states by visual inspection.

Figure 2.

Observed intensity measurements at three SNPs in a schizophrenia study (Shi et al. 2009). The case and control samples are shown in red and black, respectively. The upper panel pertains to SNP rs10771631, whose measurements suggest copy number deletion and differential measurement errors between cases and controls; the middle panel pertains to SNP rs10847679, whose measurements suggest some differences in the ASCN frequency; the lower panel pertains to SNP rs11259762, whose measurements suggest copy number duplication.

Figure 3.

Type I error for testing no genetic effect with CNV deletion (upper panel) and CNV duplication (lower panel) in the presence of differential measurement errors: in (a) and (c), the mean of measurements in cases differs from that of controls by μ ; in (b) and (d), the variance of measurements in cases is σ^2 times that of controls. The green line indicates the nominal significance level of 1%.

Figure 4. Power at the 1% nominal significance level for (a) testing the effect of the total copy number with the LRR and BAF measurements and (b) testing the gene-environment interaction with the LRR measurements only.

Figure 5.

Q-Q plots for the two-degrees-of-freedom tests at SNP loci for (a) the proposed method (genomic control $\lambda = 0.96$), (b) imputation-C ($\lambda = 1.53$), and (c) imputation-S ($\lambda = 1.27$). The black points pertain to 250 SNPs that have CNV deletions only and produce converged results. The gray dashed lines indicate 95% confidence intervals. The green straight lines represent the null distribution.

Table 1
Simulation results for studying the gene-environment interaction under gene- environment dependence/independence

	β_3	Prospective likelihood (4) (allowing G-E dependence)					Retrospective likelihood (5) (assuming G-E independence)				
		Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power
G-E dependence	.0	-.002	.087	.086	.989	.011	.099	.059	.059	.822	.178
	.1	-.002	.088	.087	.988	.085	.099	.061	.060	.837	.765
	.2	.001	.087	.087	.992	.392	.096	.061	.062	.854	.986
	.3	.000	.092	.090	.990	.774	.093	.064	.064	.874	1.000
G-E independence	.0	-.001	.085	.084	.990	.010	-.001	.059	.059	.990	.010
	.1	.000	.087	.085	.992	.081	-.001	.059	.060	.990	.177
	.2	.000	.087	.087	.991	.394	-.002	.060	.061	.989	.756
	.3	-.002	.090	.090	.991	.774	-.006	.062	.062	.991	.986

NOTE: Bias and SE are the bias and standard error of $\hat{\beta}_3$. SEE is the mean of the standard error estimator. CP is the coverage probability of the 99% confidence interval. Power is the power for testing $H_0 : \beta_3 = 0$ at the nominal significance level of 1% and pertains to the type I error when $\beta_3 = 0$.

Table 2
Simulation results for studying the genetic effect when there are no differential measurement errors

β_0	Proposed					Imputation-C					Imputation-S				
	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power
.00	.000	.067	.066	.990	.010	.000	.064	.064	.991	.009	.000	.075	.064	.971	.029
.14	.000	.067	.066	.990	.331	-.007	.065	.064	.989	.320	.003	.076	.064	.968	.388
.18	.000	.067	.066	.990	.558	-.009	.065	.064	.990	.544	.004	.075	.064	.971	.604
.22	.001	.066	.066	.992	.776	-.011	.064	.064	.988	.760	.006	.075	.064	.973	.800
.26	.001	.066	.066	.990	.916	-.013	.064	.064	.989	.905	.006	.074	.064	.973	.916
.30	.000	.067	.066	.991	.977	-.016	.064	.064	.988	.973	.007	.075	.064	.971	.973
.40	.001	.068	.066	.989	1.00	-.022	.065	.064	.984	.999	.010	.077	.064	.965	.999
.60	-.001	.067	.067	.991	1.00	-.036	.065	.065	.977	1.00	.014	.078	.066	.972	1.00

NOTE: See the Note to Table 1.

Table 3
Simulation results for studying the effect of the total copy number

β_0	Proposed					Barnes et al.				
	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power
.0	.000	.093	.091	.988	.012	.001	.128	.125	.992	.008
.1	-.001	.092	.093	.993	.066	.002	.127	.129	.993	.029
.2	-.001	.095	.095	.992	.308	.002	.134	.134	.991	.128
.3	.000	.098	.098	.991	.690	.003	.138	.139	.994	.330
.4	-.005	.110	.109	.990	.922	.004	.147	.146	.987	.576

NOTE: See the Note to Table 1.

Table 4

P-values for association tests at three SNP loci

SNP ID	Chr	Position	Type	Freq	Method	$\beta_1 = \beta_2 = 0$	$\beta_1 = 0$	$\beta_2 = 0$
rs10771631	12	30132637	deletion	.119	Proposed	1.8×10^{-2}	5.5×10^{-3}	8.9×10^{-1}
					Imputation-C	1.2×10^{-5}	9.7×10^{-6}	9.6×10^{-1}
					Imputation-S	9.3×10^{-7}	1.1×10^{-6}	8.6×10^{-1}
rs10847679	12	127797266	deletion	.036	Proposed	3.0×10^{-3}	4.6×10^{-3}	4.5×10^{-2}
					Imputation-C	1.6×10^{-3}	1.4×10^{-3}	8.0×10^{-2}
					Imputation-S	1.5×10^{-4}	1.1×10^{-4}	7.4×10^{-2}
rs11259762	10	47110763	duplication	.039	Proposed	5.1×10^{-1}	2.5×10^{-1}	8.6×10^{-1}
					Imputation-C	9.0×10^{-2}	3.2×10^{-2}	7.3×10^{-1}
					Imputation-S	1.5×10^{-1}	5.4×10^{-2}	7.6×10^{-1}

NOTE: “Chr” and “Position” are the chromosome number and physical position. “Freq” is the frequency of the allele with deletion or duplication.