TESTING FOR SPATIAL CORRELATION AND SEMIPARAMETRIC SPATIAL MODELING OF BINARY OUTCOMES WITH APPLICATION TO ABERRANT CRYPT FOCI IN COLON CARCINOGENESIS EXPERIMENTS

A Dissertation

by

TATIYANA V. APANASOVICH

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2004

Major Subject: Statistics

TESTING FOR SPATIAL CORRELATION AND SEMIPARAMETRIC SPATIAL MODELING OF BINARY OUTCOMES WITH APPLICATION TO ABERRANT CRYPT FOCI IN COLON CARCINOGENESIS EXPERIMENTS

A Dissertation

by

TATIYANA V. APANASOVICH

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Raymond J. Carroll (Chair of Committee)	Bani K. Mallick (Member)
Michael Sherman (Member)	Nancy D. Turner (Member)
Naisyin Wang (Member)	Michael Longnecker (Interim Head of Department)

August 2004

Major Subject: Statistics

ABSTRACT

Testing for Spatial Correlation and Semiparametric Spatial Modeling of Binary Outcomes with Application to Aberrant Crypt Foci in Colon Carcinogenesis Experiments. (August 2004)

Tatiyana V. Apanasovich, Dipl., Belarusian State University

Chair of Advisory Committee: Dr. Raymond J. Carroll

In an experiment to understand colon carcinogenesis, all animals were exposed to a carcinogen while half the animals were also exposed to radiation. Spatially, we measured the existence of aberrant crypt foci (ACF), namely morphologically changed colonic crypts that are known to be precursors of colon cancer development. The biological question of interest is whether the locations of these ACFs are spatially correlated: if so, this indicates that damage to the colon due to carcinogens and radiation is localized. Statistically, the data take the form of binary outcomes (corresponding to the existence of an ACF) on a regular grid. We develop score–type methods based upon the Matern and conditionally autoregression (CAR) correlation models to test for the spatial correlation in such data, while allowing for nonstationarity. Because of a technical peculiarity of the score–type test, we also develop robust versions of the method. The methods are compared to a generalization of Moran's test for continuous outcomes, and are shown via simulation to have the potential for increased power. When applied to our data, the methods indicate the existence of spatial correlation, and hence indicate localization of damage. Assuming that there are correlations in the locations of the ACF, the questions are how great are these correlations, and whether the correlation structures differ when an animal is exposed to radiation. To understand the extent of the correlation, we cast the problem as a spatial binary regression, where binary responses arise from an underlying Gaussian

latent process. We model these marginal probabilities of ACF semiparametrically, using fixed-knot penalized regression splines and single-index models. We fit the models using pairwise pseudolikelihood methods. Assuming that the underlying latent process is strongly mixing, known to be the case for many Gaussian processes, we prove asymptotic normality of the methods. The penalized regression splines have penalty parameters that must converge to zero asymptotically: we derive rates for these parameters that do and do not lead to an asymptotic bias, and we derive the optimal rate of convergence for them. Finally, we apply the methods to the data from our experiment.

To my Parents, Vladimir and Tamara Apanasovich

ACKNOWLEDGMENTS

As I type this page, I am overwhelmed for two reasons. First is the very fact that this marks the culmination of the dissertation process and my entire graduate school career. Second, I realize that I would not be at this stage had it not been for many individuals who have guided and supported me through these five years. I am grateful to my advisor, Raymond J. Carroll, for his generous criticism and encouragement of my work from beginning to end. A mentor par excellence, he has challenged and supported me and my work in innumerable ways. I would like to acknowledge the special contribution made by the members of my dissertation committee, whose comments and suggestions inspired and directed this research. I wish to express my deep appreciation to Naisyin Wang for offering encouragement, practical advice and wisdom when it was needed most. Bani K. Mallick deserves all the thanks I can convey for his inspirational wisdom and boundless enthusiasm. I owe a special debt to Michael Sherman for introducing me to the exciting field of Spatial Statistics. A very big thank you to Nancy D. Turner, for her patience in explaining the biological meaning of the experiments and for facilitating thinking about the practical issues of the dissertation.

For lessons both in and out of the classroom, I owe a profound debt to the following people: James Calvin and Michael Longnecker.

Special thanks are also due to Cliff Spiegelman, who retained me as a research assistant for several semesters, for his inspiration and invaluable experience. A big, heartfelt thank you to Dr. Dahm for his effort to recruit me for this department five years ago and for his helpful advice throughout those years.

I would like to thank Simon Sheather, who collaborated with me on my first paper and offered me useful comments and suggestions. Special thanks Joanne R. Lupton,

Natasa Popovic, and Robert S. Chapkin, who helped me gain a better understanding of the mechanisms of action in the relationship between nutrition and cancer and learn how to function as a true collaborator in teams of biologists. I am fortunate to have found wonderful colleagues and friends among students at Texas A&M. My life and work in graduate school have been richer for it. I would also like to thank my parents and my sister Natasha. Without their love, help and support my college and graduate education would have never been possible. Finally, I want to express my gratitude to Alexei Milkov for his constant support through the many hard times when my confidence and enthusiasm had waned.

This work has been supported by grants from the National Cancer Institute (CA57030, CA61750, CA82907), NASA (NPFR00202) and by the Texas A&M Center for Environmental and Rural Health through a grant from the National Institute of Environmental Health Sciences (P30-ES09106) and by the Australian Graduate School of Management.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

CHAPTER I

INTRODUCTION

The first part of the dissertation is concerned with testing for spatial correlation when the outcomes are binary. The problem arises naturally from an important question in colon carcinogenesis. In our experiments, the colon can be thought of as a cylindrical tube, which is cut lengthwise into two pieces. One piece is used for other experiments, while the other is laid out flat onto a slide, see Figure 1. Animals are exposed to a carcinogen, with half of them also exposed to radiation. They are then sacrificed, and images of the colon are obtained by various staining devices. A typical image is given in Figure 2: a color version of this is given at http://stat.tamu.edu/∼carroll/techreports.html. Here we see three types of structures:

1. The grayish region is lymphatic tissue, called Peyer's Patches.

2. The small white dots are normal colonic crypts, whose function is to produce cells that line the colon. More details accessible by a statistical audience on the role of colonic crypts are given in Morris, et al. (2001, 2002, 2003).

3. The larger dark and distended regions are aberrant crypt foci, or ACF for short, which are crypts that have been changed morphologically by the carcinogen and radiation. For technical reasons, it is not possible to determine accurately the existence of an ACF within lymphatic tissue (Peyer's Patches). See Bird (1995) and Bird and Good (2000) for the importance of ACF in colon carcinogenesis.

The journal model is *Biometrics*.

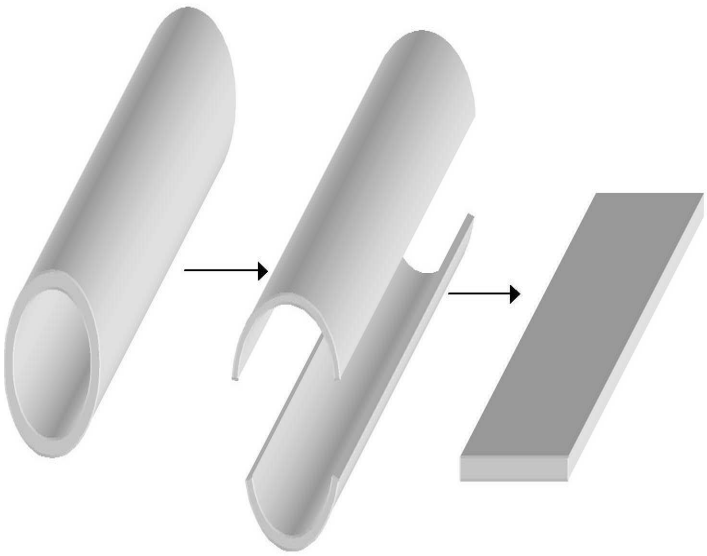

Figure 1. A drawing showing the process of laying the colon onto a slide.

Figure 2. A colon laid lengthwise, showing a Peyer's Patch (gray region), normal colon crypts (white dots) and aberrant crypt foci (dark distended shapes).

Because the Peyer's Patches are physically different tissue, we believe that it is only sensible to treat the responses in that tissue as missing complete at random. Our interest is in the aberrant crypt foci, which we denote by ACF. These are precursors to colon cancer, and hence almost everything about them are of biological relevance.

The data are clearly naturally spatial. By any measure, they are also nonstationary, as the proximal (front) and distal (back) regions of the colon behave far differently in terms of the likelihood of ACF formation. It is not feasible in practice to measure the locations of ACF, so we formed a rectangular grid of locations and recorded (by hand) the existence of an ACF within each location, see Figure 3 for an illustration. Thus the data available to us are the grid of locations along with the binary indicator of an ACF.

Here we consider a particular problem, namely testing whether the existence of an ACF at one location is predictive of an ACF at neighboring locations. Hence, we want to test for spatial dependence, using the binary outcome of the existence of an ACF. Such spatial dependence, if it exists, is interesting because it suggests that damage to the colon is localized regionally, and thus that there may be areas in which greater levels of damage in response to an insult could lead to focused areas of inflammatory responses, or an alteration in the release of signaling molecules that could then affect the regulation of homeostatic mechanisms in colonocytes in adjacent crypts. This localization may help explain why tumors develop from particular ACF, but not from all ACF formed in response to a carcinogen insult.

Thus, testing for the spatial dependence is of biological interest in itself, and as such it is not merely testing for a nuisance parameter.

One way to test for such spatial dependence is to build a spatial regression model that includes independence as a special case, and then to test for this special case.

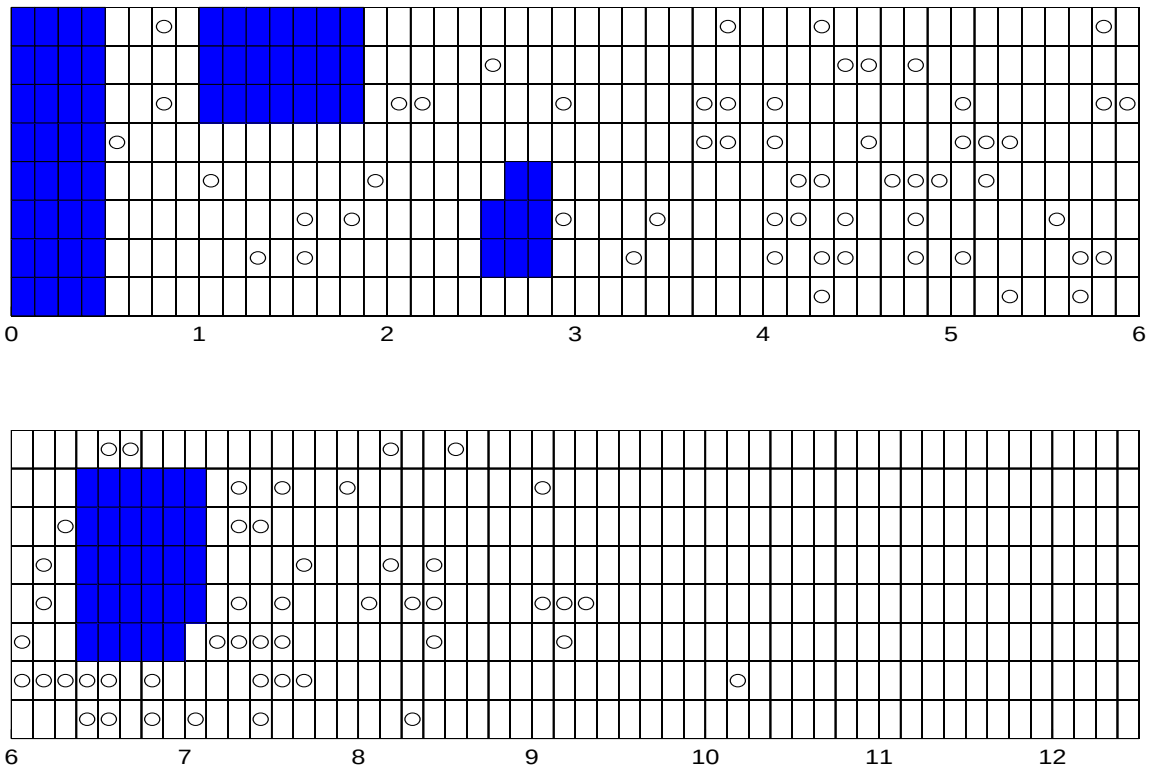

Figure 3. A gridded plot of a rat #263. Dots are coded for ACFs indicators and shaded areas are Peyer's Patches.

We instead develop simpler methods that avoid the need to fit any particular spatial model, while still allowing for the nonstationarity in the mean that is inherent in our problem. Hence the method developed has the potential to be widely applicable in practice.

One of the methods we develop is the binary version of Moran's test (Moran, 1948) that allows for spatial nonstationarity. However, our primary focus is on a test motivated by score testing ideas, and robust versions of this test that ensure that one or two neighboring pairs of ACFs will not in themselves lead to a declaration of spatial dependence.

An outline of the second chapter is as follows. In Section B, we describe Moran's test and derive the score–type test, which is based upon the Matern and conditionally autoregression (CAR) correlation models. In Section B we describe in detail the main robustness issue with the score–type test and derive robust alternatives to this test. Section C describes a set of simulations that suggest that the score test and its robust modifications can be more powerful than Moran's test, as well as having test level closer to the nominal. Section D returns to the Aberrant Crypt Foci data in detail and shows some evidence of spatial correlation, and hence localization of damage to the colon. The last Chapter has concluding remarks. All technical details are given in an Appendix A.

The main purpose of the second part of the dissertation is to develop methods for semiparametric regression using regression splines in correlated binary data problems, especially longitudinal and spatial data. Along with the methods, we develop an asymptotic theory that encompasses smoothing parameter estimation.

Having shown dependence, two questions immediately arise: the extent of the dependence and the nature of the rate of ACF formation depending on the location within the colon. Our hope in this experiment is to identify regions of high ACF

formation, and then to see whether regions of high ACF formation are also the regions of high tumor formation.

To solve this problem, we propose a binary mixed model that incorporates fairly general forms of dependence. The fixed effects structure we study includes partially linear models, single index model and 2-function additive models as special cases, using the technology of fixed knot regression splines with penalties (Ruppert, et al., 2003).

An outline of the third chapter is as follows. In Section A we describe the general class of models to which our results apply: these include a combination of partially linear and single-index models. Section B describe the basic methods of estimation, while Section D gives our algorithms for smoothing parameter estimation. The statement of asymptotic results is given in Section C, while standard error estimation is discussed in Section E. A small simulation study is presented in Section F, followed by the analysis of the ACF experiment in Section G. Discussion and extensions are given in the last Chapter. All proofs are Appendix B.

CHAPTER II

TESTING FOR SPATIAL CORRELATION IN NONSTATIONARY BINARY DATA

A. Methods

Generalized linear mixed models are widely used models for spatially dependent binary data (Breslow and Clayton, 1993; Diggle et. al 1997). These models are convenient for modeling the dependence of a response variable, Y_i, measured at $i = 1, ..., n$ sites, as well as on measured covariates, X_i. We use a multivariate probit model to model spatial dependence and nonstationarity. Let $\mathrm{I}(\cdot)$ be the indicator function. Let the ϵ_i for $i = 1, ..., n$ be independent and normally distributed with mean 0 and variance 1. Let λ_i denote random effects responsible for possible spatial dependence. For a parameter ρ and a correlation matrix $\Omega(\rho)$, the λs are assumed to be normally distributed with mean 0 and covariance matrix $\sigma^2_\lambda \Omega(\rho)$. Let μ_i be systematic effects incorporating nonstationarity. Then the multivariate probit model is defined as $Y_i = \mathrm{I}(\mu_i + \lambda_i + \epsilon_i > 0)$, so that

$$\mathrm{pr}(Y_i = 1 \mid \lambda_i, \mu_i) \quad = \quad \Phi(\mu_i + \lambda_i), \tag{2.1}$$

where $\Phi(\bullet)$ is the univariate standard normal distribution function. We are interested in assessing whether $\rho = 0$, in which case $\Omega(0)$ is the identity matrix.

The rest of this section is taken up with defining the two methods we use.

1. Moran's Test

Moran's test for spatial dependence (Moran, 1948) was developed for stationary data. For stationary data $\mu_i = \mu(X_i) \equiv \mu$, and the test is as follows. Let Z^{vec} be the vector

of observations Y minus their sample mean. Let W^{mat} be an $n \times n$ matrix with (i, j) element equal to 1 if sites (i, j), $i \neq j$ are neighbors and equal 0 otherwise. Moran's test statistic is (up to a constant of proportionality) $(Z^{vec})^{\mathrm{T}} W^{mat} Z^{vec} / (Z^{vec})^{\mathrm{T}} Z^{vec}$. Note that Moran's test statistic takes on the classic form of any autocorrelation coefficient: the numerator term is a measure of covariance and the denominator term is a measure of variance. Its values are compared to a non-trivial expression, see Cliff and Ord (1981, Chapter 1, pp. 19-21).

For nonstationary numerical data, Moran's test is usually modified (Cliff and Ord, 1981) by subtracting predicted values from the observations rather than the mean.

For our case of nonstationary binary data, we modify Moran's test in the usual way, namely by letting Z^{vec} be the vector of standardized residuals from an ordinary Probit regression of the Ys on the Xs: $(Y_i - \widehat{Y}_i)/\{\widehat{Y}_i(1 - \widehat{Y}_i)\}^{1/2}$.

2. Score Test

Rao's score statistic (Rao, 1973) is a standard tool for carrying out hypothesis testing. In many situations it has the advantage over likelihood ratio and Wald tests because all calculations are carried out under the hypothesis, except the derivation of the test statistic itself. In our context of a multivariate probit model, it does not appear possible to derive such an explicit formula for the score test for an arbitrary correlation function.

We instead take a different approach, one that yields a readily computed test statistic. Our idea is to look at pairs of observations, and derive a score test statistic for correlation using such pairs when the correlation is of the Matern class (Stein, 1999). We will then combine this test statistic over many pairs. As we show in the Appendix, it turns out that the resulting test is the same as the score test for a

particular version of the conditionally autoregressive (CAR) correlation model (Besag, 1974; Richardson, at al., 1992).

a. Score Statistic for Pairs

We first compute the joint probability distribution of any two binary responses Y_i and Y_j. Let $\Phi(\bullet)$ and $\phi(\bullet)$ be the univariate standard normal distribution and density functions, respectively. Let $\Phi_2(\mu_1, \mu_2, \rho)$ be the bivariate standard normal probability of being below μ_1 and μ_2 when the correlation is ρ. Define $\mu_i^* = \mu_i/(1+\sigma_\lambda^2)^{1/2}$. Then

$$\begin{aligned} \text{pr}(Y_i = 1|\mu_i) &= \Phi(\mu_i^*); \qquad (2.2) \\ \text{pr}(Y_i = 1, Y_j = 1|\mu_i, \mu_j) &= \Phi_2\{\mu_i^*, \mu_j^*, \sigma_\lambda^2 \Omega_{ij}(\rho)/(1+\sigma_\lambda^2)\}. \end{aligned}$$

We will calculate the score–type test based on $k = 1, ..., N$ pairs. Consider the kth pair $(Y_{1k} = i, Y_{2k} = j)$, where we write $\text{pr}(Y_{1k} = 1) = \Phi(\mu_{1k}^*)$ and $\text{pr}(Y_{2k} = 1) = \Phi(\mu_{2k}^*)$. Also define $\text{pr}(Y_{1k} = i, Y_{2k} = j) = \pi_{ijk}(\rho, \sigma_\lambda^2) = \pi_{ijk}$, so that $\text{pr}(Y_{1k} = i) = \pi_{i\cdot k}$ and $\text{pr}(Y_{2k} = j) = \pi_{\cdot jk}$, where the "dots" indicate summation. A useful fact is that

$$\pi_{11k}(\rho = 0, \sigma_\lambda^2) = \Phi(\mu_{1k}^*)\Phi(\mu_{2k}^*). \qquad (2.3)$$

If we define $Z_{ijk} = \text{I}(Y_{1k} = i, Y_{2k} = j)$, then the loglikelihood is

$$\text{logL}(\rho) = \sum_k \{Z_{00k}\log(\pi_{00k}) + Z_{01k}\log(\pi_{01k}) + Z_{10k}\log(\pi_{10k}) + Z_{11k}\log(\pi_{11k})\}. \qquad (2.4)$$

Formal differentiation of (2.4) and evaluated at the null hypothesis $\rho = 0$ would yield the essential part of the score statistic. Let d_k be the Euclidean distance between the members of the kth pair. Recent literature (e.g. Stein, 1999, p.31–33) advocates the use of the Matern family, for which the covariance functions have general form

$$C_M(d_k) = \frac{\sigma_S^2}{2^{\nu-1}\Gamma(\nu)}(d_k/\rho)^\nu K_\nu(d_k/\rho), \qquad \sigma_S^2, \rho, \nu > 0, \qquad (2.5)$$

where K_ν is the modified Bessel function of order ν, which do not have a closed form for general ν. If $\nu = m + \frac{1}{2}$ for $m = 0, 1, 2, ...$, then (2.5) has a simple form. However, as we show in the Appendix, there is a difficulty with this approach. When we use as the correlation function a member of the Matern class with $\nu = m + \frac{1}{2}$, the score evaluated at $\rho = 0$ is identically 0, so that nothing useful results. As we show in the Appendix, our approach is to focus only on those pairs that are exactly the same distance apart, in which case the score becomes a non–trivial statistic times a common constant that equals 0 when $\rho = 0$. Removing this common constant leads to a score equal to

$$\mathcal{G}_k(\mu^*_{1k}, \mu^*_{2k}) = \frac{(Y_{1k} - \pi_{1\cdot k})(Y_{2k} - \pi_{\cdot 1k})\phi(\mu^*_{1k})\phi(\mu^*_{2k})}{\pi_{1\cdot k}(1 - \pi_{1\cdot k})\pi_{\cdot 1k}(1 - \pi_{\cdot 1k})}. \tag{2.6}$$

In practice, we implement our score–type test as follows. Recall model (3.1), and let the μs depend on covariates X and a parameter β_*, i.e., $\mu(X, \beta_*)$ with the property that for any constant c, $c\mu(X, \beta_*) = \mu(X, \beta_{**})$ for some β_{**}. As seen in (2.2), under the null hypothesis the Ys are independent and $\text{pr}(Y = 1|X) = \Phi\{\mu(X, \beta_*)/(1+\sigma^2_\lambda)^{1/2}\} = \Phi\{\mu(X, \beta)\}$, say. Thus, we can estimate β consistently under both the null and alternative models via a probit regression with probability function $\text{pr}(Y = 1|X) = \Phi\{\mu(X, \beta)\}$. Call the estimate $\widehat{\beta}$. Modify (2.6) appropriately by defining

$$\mathcal{H}_k(\beta) = \frac{[Y_{1k} - \Phi\{\mu(X_{1k}, \beta)\}][Y_{2k} - \Phi\{\mu(X_{2k}, \beta)\}]\phi\{\mu(X_{1k}, \beta)\}\phi\{\mu(X_{2k}, \beta)\}}{\Phi\{\mu(X_{1k}, \beta)\}[1 - \Phi\{\mu(X_{1k}, \beta)\}]\Phi\{\mu(X_{2k}, \beta)\}[1 - \Phi\{\mu(X_{2k}, \beta)\}]}. \tag{2.7}$$

The variance of (2.7) under the hypothesis of no spatial correlation is clearly

$$\mathcal{V}_k(\beta) = \frac{[\phi\{\mu(X_{1k}, \beta)\}\phi\{\mu(X_{2k}, \beta)\}]^2}{\Phi\{\mu(X_{1k}, \beta)\}[1 - \Phi\{\mu(X_{1k}, \beta)\}]\Phi\{\mu(X_{2k}, \beta)\}[1 - \Phi\{\mu(X_{2k}, \beta)\}]}. \tag{2.8}$$

Our test statistic then is

$$\frac{\sum_k \mathcal{H}_k(\widehat{\beta})}{\{\sum_k \mathcal{V}_k(\widehat{\beta})\}^{1/2}}. \tag{2.9}$$

We show in the Appendix that under the hypothesis of no spatial correlation, (2.9) is asymptotically standard normal and hence the hypothesis of no spatial correlation can be tested by referring (2.9) to standard normal quantiles. Notice that the terms in the test statistic's numerator are not independent, though they are uncorrelated. The method of Commenges and Jacqmin-Gadda (1997) can be used to prove asymptotic normality under the assumption that $\mu(X, \beta)$ and its first two derivatives in β are bounded. The result is asymptotic in the number of pairs, with the same scale of grid and the same spatial scale of autocorrelation.

Terms similar to (2.6) and the numerator of (2.9) were derived in a different context by le Cessie and van Houwelingen (1994). They considered the case of classical clustered and not spatial data, and this context is crucial because in our problem, the number of pairs within each animal/colon is large. Even taking this difference of context into account, there still remain important differences with our work. They considered the case, in effect, that the correlation matrix $\Omega(\rho)$ has common correlation for all elements, something not likely to hold for spatial data. Their sum in (2.9) would thus be over all pairs, and not just pairs of neighbors. In addition, they required multiple clusters (animals), and because of their different context were not led (a) to notice the problem raised above with straightforward use of the Matern class; and (b) to show that under the null hypothesis of no spatial correlation, for a single animal and a large number of locations, the denominator of (2.9) is a consistent estimator of the standard deviation of the numerator under the hypothesis of no spatial correlation, taking into account the estimation of β.

Also somewhat similar to our test is work of Jacqmin-Gadda, et al. (1997), with their version of (2.7) having elements of the form $[Y_{1k} - \Phi\{\mu(X_{1k}, \beta)\}][Y_{2k} - \Phi\{\mu(X_{2k}, \beta)\}]w(X_{1k}, X_{2k})$ across all pairs (not just necessarily neighbors) and for an arbitrary function $w(\bullet)$. Their motivation and actual test statistics are however very

different: in place of our (3.1) they start from a logistic family with a correlation model for the λ-terms that is fixed in advance. In contrast, our work and that of le Cessie and van Houwelingen is based on somewhat more standard spatial correlation models involving a free parameter: Matern and CAR in our case, and equicorrelated for le Cessie and van Houwelingen.

3. Selecting the Pairs

There are many ways to organize all observations into pairs, bearing in mind that our score–type test is based on the idea that observations in all pairs should be the same distance apart. What we do is the following. Each observation is paired with its closest neighbors in any vertical and horizontal direction, so that each observation will make as many pairs as it has neighbors. Hence, interior observations will have four pairs, ones on edges will have three pairs and ones in corners will have two pairs.

4. Conditional Autoregressive Models (CAR)

A simple version of the conditionally autoregressive (CAR) correlation model (Besag, 1974; Richardson, et al., 1992) is that the λs have covariance matrix $\sigma^2_\lambda(I - \rho C)^{-1}$, where C is chosen to be a neighborhood matrix whose (i, j)th element is equal to 1 if region i and region j $(i \neq j)$ are neighbors and I is an identity matrix of appropriate dimension. As we show in the Appendix, Section **??**, our test (2.9) is the same as the score test for this model, and in this regard is more general than simply the Matern class.

B. Robust Score Tests

If the event rates are rare, then one would not expect to have two neighboring pairs of observations for which both Ys equal 1, unless the correlations are reasonably high. Because of this, one would expect that any score-type test would have the property that when the event rates are small, a pair of neighboring Ys equal 1 would lead to the rejection of the hypothesis of no spatial correlation. Our test does indeed have this property. In fact, if $Y_{1k} = Y_{2k} = 1$, then (2.7) becomes $\phi\{\mu(X_{1k},\beta)\}\phi\{\mu(X_{2k},\beta)\}/[\Phi\{\mu(X_{1k},\beta)\}\Phi\{\mu(X_{2k},\beta)\}]$, which is unbounded as $\mu(X_{1k},\beta) \to -\infty$ and $\mu(X_{2k},\beta) \to -\infty$. On the other hand, the variance contribution in (2.8) is bounded in such a circumstance, in fact converges to 0. This means that with such a single pair, the test statistic (2.9) converges to ∞, as intuition suggests.

The above fact may be looked upon as a strength of the score–type test, but it may also be a flaw. In a particular data set, there may be little evidence of spatial correlation except for a single pair, and this would make one wary of claiming such a correlation.

In this section, we propose a simple modification of our score–type test that limits the influence of any one pair on the value of the score–type test, while still allowing for considerable power. The method is based on ideas from robustness theory.

To develop this method, rewrite (2.7) as follows:

$$\begin{aligned} \mathcal{H}_k(\beta) &= [Y_{1k} - \Phi\{\mu(X_{1k},\beta)\}][Y_{2k} - \Phi\{\mu(X_{2k},\beta)\}]R(X_{1k},X_{2k},\beta); \quad (2.10) \\ R(X_{1k},X_{2k},\beta) &= \frac{\phi\{\mu(X_{1k},\beta)\}\phi\{\mu(X_{2k},\beta)\}}{\Phi\{\mu(X_{1k},\beta)\}[1-\Phi\{\mu(X_{1k},\beta)\}]\Phi\{\mu(X_{2k},\beta)\}[1-\Phi\{\mu(X_{2k},\beta)\}]}. \end{aligned}$$

It is the function $R(\bullet)$ that is unbounded, and hence in the terminology of robustness, (2.10) is a statistic with unbounded influence. Note that $R(\bullet)$ does not depend on

the responses, and in this respect acts in a fashion similar to that of a design matrix in linear regression.

Methods to bound the influence of "design" points in logistic and linear regression have been investigated by Carroll and Pederson (1993) and by Simpson, et al. (1992), respectively. The idea is to redefine the test statistic (2.10) as

$$\mathcal{H}_{k,robust}(\beta) = [Y_{1k} - \Phi\{\mu(X_{1k},\beta)\}][Y_{2k} - \Phi\{\mu(X_{2k},\beta)\}]H\left\{R(X_{1k},X_{2k},\beta)\right\},$$

for an arbitrary function $H(\bullet)$, and to redefine its variance as

$$\begin{aligned}\mathcal{V}_{k,robust}(\beta) &= \Phi\{\mu(X_{1k},\beta)\}[1-\Phi\{\mu(X_{1k},\beta)\}] \\ &\quad \times\Phi\{\mu(X_{2k},\beta)\}[1-\Phi\{\mu(X_{2k},\beta)\}]H^2\left\{R(X_{1k},X_{2k},\beta)\right\}.\end{aligned}$$

We now define two classes of weight functions. The first follows Carroll and Pederson (1993). Let σ_R be thge median of the terms $R(X_{1k},X_{2k},\beta)$. Define $L_k = R(X_{1k},X_{2k},\beta)/\sigma_R$ and for a constant b_{cp}, define $H_{cp}\{R(X_{1k},X_{2k},\beta)\} = R(X_{1k},X_{2k},\beta)$ $\{1-(L_k/b_{cp})^2\}^3 I(L_k \le b_{cp})$. If $b_{cp}=\infty$, we of course get the score–type test, while smaller values of b_{cp} bound the influence. We experimented with some simulated data, and finally choose $b_{cp}=3$.

The method of Simpson, et al. is similar, namely $H_{simpson}\{R(X_{1k},X_{2k},\beta)\} = R(X_{1k},X_{2k},\beta)\min\{1,(b_{simpson}/L_k)^{\alpha/2}\}$. We took $(b_{simpson}=1,\alpha=1)$ and $(b_{simpson}=2,\alpha=2)$.

C. Simulations

We performed simulations under two scenarios with the number of replications equal to 1000. In both cases, we took the test level to be 0.05.

In Scenario #1, we took data from a rat labeled as #263 in our experiment. Let

X be the horizontal distance from the distal part of the colon, let Z be the vertical distance, and let D be the Euclidean distance to the nearest Peyer's Patch. We fit a probit model to these data with probability function

$$\Phi(\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 Z + \beta_4 D). \tag{2.11}$$

The original grid was of size 100×8. For the simulations we used half of that grid, 50×8. Data were generated from the ordinary probit fit to this model, with $\beta_0 = -1.83$, $\beta_1 = 6.96$, $\beta_2 = -7.34$, $\beta_3 = -3.12$, $\beta_4 = -3.46$. We generated correlated data with $\sigma^2_\lambda = 1$ via the Matern correlation function with index 3/2: $\Omega_{ij}(\rho) = \text{corr}(\lambda_i, \lambda_j) = \exp(-d_{ij}/\rho)\,(1 + d_{ij}/\rho)$.

In Scenario #2, data were generated from the model with the following probability function $\Phi(\beta_0+\beta_1 X+\beta_2 X^2)$ with $\beta_0 = -4.5$, $\beta_1 = 12.03$, $\beta_2 = -8$, $X \in [0.0186, 1]$. The grid was taken to be of size 54×8. We generated correlated data with $\sigma^2_\lambda = 1$ via the Matern correlation function with index 3/2.

Since we select pairs that are of the same distance apart, all Ω_{ij} are equal to, say, Ω. Define $\psi = \sigma^2_\lambda/(1+\sigma^2_\lambda)\Omega$, see the Appendix. For power comparison purposes, we vary ψ.

In addition to the test statistics described previously, we also computed the level and power for the test statistic based on (2.10) but with the function $R(\bullet) \equiv 1$. The motivation for this comes from logistic regression. If in (2.10) the normal distribution function $\Phi(\bullet)$ and its density $\phi(\bullet)$ were replaced by the logistic distribution and density functions, then $R(\bullet) \equiv 1$. Given estimates of β the test would formally have bounded influence, and it is one of the tests evaluated in the simulation study done by Jacqmin-Gadda, et al. (1997). The results are given in Table 1.

Table 1.

Results of the simulations. Comparison of Test performance under different scenarios

Simulation Scenario #1							
$\psi =$	0.00	0.10	0.15	0.20	0.25	0.30	0.40
score	0.05	0.13	0.28	0.46	0.63	0.73	0.82
moran	0.06	0.14	0.26	0.42	0.60	0.66	0.75
adj.moran	0.05	0.11	0.20	0.36	0.55	0.61	0.72
scorecp	0.05	0.13	0.28	0.46	0.63	0.73	0.82
scores1	0.05	0.13	0.28	0.46	0.63	0.73	0.82
scores2	0.05	0.13	0.28	0.46	0.63	0.73	0.82
scoreJG	0.05	0.13	0.28	0.46	0.63	0.73	0.82
Simulation Scenario #2							
$\psi =$	0.00	0.10	0.15	0.20	0.25	0.30	0.40
score	0.05	0.22	0.48	0.73	0.89	0.95	0.98
moran	0.07	0.24	0.45	0.68	0.83	0.90	0.91
adj. moran	0.05	0.21	0.42	0.65	0.81	0.88	0.90
scorecp	0.05	0.22	0.48	0.73	0.89	0.95	0.98
scores1	0.05	0.22	0.48	0.73	0.89	0.95	0.98
scores2	0.05	0.22	0.48	0.73	0.89	0.95	0.98
scoresJG	0.05	0.22	0.48	0.73	0.89	0.95	0.98

The number of replications is 1000. The parameter $\psi = \sigma_\lambda^2/(1+\sigma_\lambda^2)\Omega$ indicates the strength of the spatial correlation, with $\psi = 0$ being the case of no spatial dependence. Hence, the column $\psi = 0.00$ is the level of the test, while the other columns are the powers. Scenario #1 and Scenario #2 are described in the text. The tests are as follows. "score" refers to the usual score test, "moran" to Moran's test, "scorecp" to the robust score test with Carroll and Pederson's weight function, "scores1" to Simpson's weight function with $b = 3$, $\alpha = 1$, "scores2" to Simpson's weight function with $b = 4$ and $\alpha = 2$ and "scoreJG" to Score test similar to Jacqmin-Gadda, et al. (1997). "adj. moran" refers to Moran's test adjusted to have level exactly 0.05.

We observed in a variety of simulations that Moran's test had a tendency to be anticonservative and fail to maintain the level of 0.05. Because of this, we also display results when Moran's test was adjusted to have exact level 0.05. This was done by running the null case of no spatial correlation many times, computing Moran's test statistic, and then choosing as the rejection point that value that gave exact null level 0.05. For both scenarios all the score tests came reasonably close to maintaining the nominal level. In addition, the power of the robust score tests and the Original Score Test are nearly the same. In both scenarios score tests have greater power than Moran's Test, when the latter was adjusted to have correct level. The test discussed above in which $R(\bullet) \equiv 1$ does quite well in terms of power in these simulations. However, we have done other simulations where we do see some loss of power with this choice.

D. Aberrant Crypt Foci (ACF) Experiment

The introduction describes the ACF experiment, but here we make a few additional remarks about the data collection. The typical rat colon was approximately 10cm-12cm long when laid out on a slide. This is far larger than can be read in one go from a microscope. Instead, what was done was to first start with a piece of paper, somewhat like that given in Figure 2, but without the grid lines superimposed. We then simply physically moved the slide, horizontally and vertically, starting from the proximal part of the colon, through the microscope, noting approximately how far along we were in physical (slide) distance. As we observed ACFs and Peyer's Patches, we made small notations in pencil on the piece of paper.

The image in Figure 1 is approximately 1/5–1/6 square centimeters, and thus takes up approximately 9 grid boxes. This, by the way, is the only image that

was recorded, and then primarily for the purpose in this paper of illustrating ACFs. One can see multiple (7) ACFs, along with a small Peyer's Patch: in other sections everything seen in the microscope was a Peyer's Patch, as illustrated in Figure 2. What was recorded then was the approximate location of this square image, as well as the existence of ACFs and Peyer's Patches.

Since the work was done manually, the locations of the ACFs and Peyer's Patches as marked on the paper are not exact. The first and last authors observed the process numerous times, and on this basis and in collaboration with our colleagues decided to use the gridding as displayed in Figure 2. We felt that any finer grid would have led to far too much misclassification of location.

All rats were exposed to a chemical carcinogen. One half of the rats were also exposed to radiation. Rats were sacrificed at 4, 6 and 8 weeks, and their colons removed and assayed. There were thus 6 rat groups, and 7 were in each group.

Using model (2.11), we first computed our score tests on an animal by animal basis. We then combined the results as follows. For rat $r = 1, ..., 7$ in rat group $g = 1, ..., 6$, each test statistic can be written as T_{rg}/S_{rg}, where T is the numerator of the test statistic and S is its denominator. Since the rats are independent, a simple way to combine the data in a group is to compute the "combined" test statistic

$$\sum_{r=1}^{7} T_{rg}/(\sum_{r=1}^{7} S_{rg}^2)^{1/2}. \tag{2.12}$$

The results are given in Tables 2–3. At 4 weeks after administration of the carcinogen, there is little evidence of strong spatial correlation, with only 1 animal in the irradiated and non–irradiated groups having evidence of correlation, and the combined test being thoroughly not statistically significant.

Table 2.

Significance levels for irradiated rats

rat	score	moran	scorecp	scores1	scores2	scoreJG
141	0.00	0.01	0.00	0.00	0.00	0.00
142	0.67	0.73	0.66	0.66	0.66	0.66
143	0.49	0.45	0.49	0.50	0.50	0.48
144	0.59	0.75	0.60	0.60	0.60	0.62
145	0.20	0.13	0.20	0.20	0.20	0.22
146	0.27	0.11	0.28	0.27	0.27	0.37
147	0.21	0.12	0.22	0.21	0.21	0.24
Combined	0.29	0.53	0.24	0.29	0.29	0.19
161	0.04	0.07	0.04	0.04	0.04	0.04
162	0.98	0.86	0.97	0.98	0.98	0.91
163	0.00	0.00	0.00	0.00	0.00	0.00
164	0.05	0.00	0.05	0.05	0.05	0.09
165	0.40	0.79	0.39	0.40	0.40	0.31
166	0.00	0.00	0.00	0.00	0.00	0.00
167	0.86	0.73	0.86	0.86	0.86	0.86
Combined	0.00	0.00	0.00	0.00	0.00	0.00
181	0.28	0.59	0.27	0.28	0.28	0.18
182	0.91	0.86	0.91	0.91	0.91	0.96
183	0.05	0.05	0.05	0.05	0.05	0.05
184	0.80	0.54	0.81	0.80	0.80	0.87
185	0.28	0.25	0.28	0.28	0.28	0.27
186	0.13	0.11	0.13	0.13	0.13	0.14
187	0.89	0.99	0.88	0.89	0.89	0.89
Combined	0.02	0.02	0.02	0.02	0.02	0.02

The tests are defined in Table 1. "Combined" is the combined test statistic. The first leading digit in the rat number indicates the animals were irradiated, the second number is the time of sacrifice and the third is the rat number in its group. The tests are as follows. "score" refers to the usual score test, "moran" to Moran's test, "scorecp" to the robust score test with Carroll and Pederson's weight function, "scores1" to Simpson's weight function with $b = 3$, $\alpha = 1$, "scores2" to Simpson's weight function with $b = 4$ and $\alpha = 2$ and "scoreJG" to Score test similar to Jacqmin-Gadda, et al. (1997). "adj. moran" refers to Moran's test adjusted to have level exactly 0.05.

Table 3.

Significance levels for non–irradiated rats

rat	score	moran	scorecp	scores1	scores2	scoreJG
241	0.64	0.39	0.64	0.64	0.64	0.71
242	0.00	0.00	0.00	0.00	0.00	0.00
243	0.52	0.72	0.46	0.51	0.52	0.44
244	0.89	0.82	0.89	0.89	0.89	0.90
245	0.78	0.40	0.77	0.78	0.78	0.92
246	0.46	0.51	0.45	0.46	0.46	0.44
247	0.51	0.45	0.51	0.51	0.51	0.56
Combined	0.38	0.12	0.38	0.38	0.38	0.47
261	0.00	0.00	0.00	0.00	0.00	0.00
262	0.02	0.00	0.04	0.02	0.02	0.12
263	0.06	0.06	0.06	0.06	0.06	0.06
264	0.25	0.23	0.25	0.25	0.25	0.32
265	0.03	0.00	0.03	0.03	0.03	0.05
266	0.52	0.67	0.53	0.52	0.52	0.44
267	0.68	0.47	0.69	0.68	0.68	0.75
Combined	0.00	0.00	0.00	0.00	0.00	0.00
281	0.03	0.00	0.03	0.03	0.03	0.06
282	0.34	0.46	0.33	0.34	0.34	0.30
283	0.67	0.31	0.66	0.67	0.67	0.83
284	0.78	0.86	0.76	0.78	0.78	0.65
285	0.81	0.61	0.82	0.81	0.81	0.91
286	0.98	0.52	0.97	0.98	0.98	0.65
287	0.00	0.00	0.00	0.00	0.00	0.00
Combined	0.00	0.00	0.00	0.00	0.00	0.00

The tests are defined in Table 1. "Combined" is the combined test statistic. The first leading digit in the rat number indicates the animals were not irradiated, the second number is the time of sacrifice and the third is the rat number in its group. The tests are as follows. "score" refers to the usual score test, "moran" to Moran's test, "scorecp" to the robust score test with Carroll and Pederson's weight function, "scores1" to Simpson's weight function with $b = 3$, $\alpha = 1$, "scores2" to Simpson's weight function with $b = 4$ and $\alpha = 2$ and "scoreJG" to Score test similar to Jacqmin-Gadda, et al. (1997). "adj. moran" refers to Moran's test adjusted to have level exactly 0.05.

However, at 6 and 8 weeks there are many more ACFs. This is perhaps not unexpected, since it takes some time after an insult via carcinogen or radiation before ACFs form. As seen in Tables 2–3, the combined tests are highly statistically significant at these time points, and 10 of the 28 individual rats show individual evidence of spatial correlation. It seems to us, then, that the evidence is fairly strong for spatial correlation at the 6–8 week time period. See McLellan, et al. (1991) for discussion of the role of time in developing ACF.

There is an interesting feature in Table 2. Although no p-value for the "18" group (irradiated rats at 8 weeks) was less than 0.05, the overall p-value using the combined test (2.12) was 0.02. This phenomenon can be explained as follows. It turns out that for these animals, the denominators S_{rg}^2 of (2.12) for rat r in group g were on average over the rats approximately equal to 14. Suppose we know that the numerators, of (2.12), T_{rg} for $r = 1, ..., 7$ are Normal$\{\mu, (14.76)^2\}$. Then each individual test has little power for testing $\mathrm{H}_0 : \mu = 0$, but the mean of the Ts has much more power. In other words, we might expect the situation for the "18" rats to be the rule, rather than the exception. Indeed, if we set μ equal to 13.14, the mean of the numerators of (2.12), the combined test (2.12) has power 66%, while the univariate tests have power 14%.

The choice of weights $R(\bullet) \equiv 1$ mentioned by Jacqmin–Gadda, et al. (1997) sometimes has quite different (higher) p–values from the score tests, with changes in statistical significance at the rat level, note for example rat 164. One might expect this to happen if the data were actually generated by a CAR model, for example.

CHAPTER III

SEMIPARAMETRIC SPATIAL MODELING OF BINARY OUTCOMES

A. Models

1. Binary Mixed Model and General Fixed Effects Structure

Our models are semiparametric forms of binary generalized linear mixed models, where a binary response variable $\{D_{ri} : i \in \mathcal{Z}^2\}$ is measured in a possibly irregular shaped set $S_{n_r} \subset \mathcal{Z}^2$ with $|S_{n_r}| = n_r$, where $|\bullet|$ is the cardinality of s set, for subjects $r = 1, ..., R$. We also measure covariates, $\widetilde{X}_{ri} = (X_{ri1}, X_{ri2})$, where X_{ri2} is scalar. In the ACF experiment, which is essentially longitudinal, X_{ri2} is the horizontal distance of the location from the distal part of the colon. For the binary model, let ϵ_{ri} be independent and normally distributed with mean 0 and variance 1. Let λ_{ri} denote random effects responsible for possible spatial dependence. For a parameter ρ and a correlation matrix $\Omega_r(\rho)$, the $\{\lambda_{ri}\}$ are assumed to be normally distributed with mean 0 and covariance matrix $\sigma^2_\lambda \Omega_r(\rho)$. Let μ_{ri} be systematic effects possibly incorporating nonstationarity. Then the model is defined as $D_{ri} = \mathrm{I}(\mu_{ri} + \lambda_{ri} + \epsilon_{ri} > 0)$, so that

$$\mathrm{pr}(D_{ri} = 1 \mid \lambda_{ri}, \mu_{ri}) \quad = \quad \Phi(\mu_{ri} + \lambda_{ri}), \tag{3.1}$$

where $\Phi(\bullet)$ is the univariate standard normal distribution function. Notice that marginally responses come from the probit model. The primary difficulty in implementing full likelihood inference for such models is that the likelihood function is intractable numerically. In the rest of this section, we describe the general fixed effects structure on the μ_{ri} (Section 2). Since our models are semiparametric in structure, we also describe (Section 3) the class of fixed-knot regression splines that form the basis for our modeling. The general random effects structure $\Omega_r(\rho)$ is introduced in

Section 4.

2. General Fixed Effects Structure

In our case, in its most general form we will allow for flexible semiparametric models both for the fixed effects μ_{ri} as well as the correlation matrix $\Omega_r(\rho)$. Specifically, for unknown functions $\Lambda_{r1}(\bullet)$ and $\Lambda_{r2}(\bullet)$, a known function $\Upsilon(\bullet)$ and an unknown parameter ζ_{r0}, we model the fixed effects structure as

$$\mu_{ri} = \Lambda_{r1}\{\zeta_{r0}^{\mathrm{T}}\Upsilon(X_{ri1})\} + \Lambda_{r2}(X_{ri2}). \tag{3.2}$$

Model (3.2) is a combination of partially linear models, additive models and single index models: (a) partially linear models obtain when $\Lambda_{r1}(\bullet)$ is known to be the identity function; (b) single index models obtain when $\Lambda_{r2}(\bullet) \equiv 0$; and (c) additive models have X_{ri1} scalar and $\zeta_{r0} = 1$ known. If $\Lambda_{r1}(\bullet)$ is unknown, then for identifiability we set $\|\zeta_{r0}\| = 1$ and insist that the first element of ζ_{r0} is positive.

3. Regression Splines and Penalization

In our work, we model an unknown function $\Lambda_{rj}(\bullet)$ as a fixed-knot regression spline, which has the representation $\Lambda_{rj}(x) = \widetilde{B}_j^{\mathrm{T}}(x)\eta_{rj0}$, where $\widetilde{B}_r^{\mathrm{T}}(x) = \{B_{r1}(x), ..., B_{rk}(x)\}$. For example, $\widetilde{B}_r(x)$ might be the Bspline basis functions (Eilers and Marx, 1996) or the truncated power series basis (Ruppert, et al., 2003) of order q with K knots $x_1, ..., x_K$ given as $\widetilde{B}_r^{\mathrm{T}}(x) = (1, x, ..., x^q, |x-x_1|_+^q, ..., |x-x_K|_+^q)$, where the subscripted plus sign is the positive part function.

Regression splines with a fixed number of knots have become an increasingly popular means of semiparametric inference, see Ruppert, et al. (2003) and references therein. Generally, not many knots are required to capture most fixed effects structures (Ruppert, 2002), and in any case for binary data in particular capturing the

type of very complex structure that cannot be captured by a low-order basis representation is unlikely to be practical. Of course some sort of smoothing is required. This is generally done either by knot selection devices to greatly lower the dimensionality, or by penalization to achieve smoothness. In this paper, we use the latter device, see Section D for details.

4. General Random Effects Structure

In this section we discuss general approach to modeling the covariance function of the spatial process $\{\lambda_{ri}\}$. We will consider a single possibility here, although we believe that our methods and theoretical techniques apply more generally.

Let $d_r(i,j)$ be the Euclidian distance between sites i and j in subject r. The simplest type of correlation structure is stationary, e.g., the Mátern family. Thus, the (i,j) element of $\Omega_r(\rho)$ can be written as $\Omega_{rij}(\rho) = \mathcal{M}\{d_r(i,j),\rho\}$, where $\mathcal{M}(\bullet)$ is a known function with unknown parameter ρ. Given the paucity of information in binary data, stationarity is clearly the default option.

There are many ways to weaken the assumption of stationarity. One approach is similar to that of Fuentes (2002), with the difference that because of the lack of information in binary data, we do not allow the variance of the latent random $\{\lambda_{ri}\}$ to vary with location. Let $K(\bullet)$ be a known symmetric density function, let h be an unknown bandwidth and define $K_h(u) = K(u/h)$. Define a set of locations as $\chi_0, ..., \chi_S$: generally, S is small. The proposed method requires that for each site there is at least one location χ_s, $s = 1, ..., S$ such that $K_h\{d_r(i,\chi_s)\} > 0$ which defines a lower bound on h. Let $\varrho_0 = 0$ and let ϱ_s, $s = 1, ..., S$ be unknown parameters. Then our model is

$$\Omega_{rij}(\varrho) \;=\; S^{-1}\sum_{s=0}^{S}\varphi\{d_r(i,\chi_s)\}\varphi\{d_r(j,\chi_s)\}\mathcal{M}\{d_r(i,j),\varrho_s\}, \tag{3.3}$$

$$\varphi\{d_r(i,\chi_s)\} \;=\; \frac{K_h\{d_r(i,\chi_s)\}}{[\max_x \sum_{t=1}^{S} K_h^2\{d_r(x,\chi_t)\}]^{1/2}}, \quad s=1,...,S,$$

and where $\varphi\{d_r(i,\chi_0)\} = [1-\sum_{s=1}^{S}\varphi^2\{d_r(i,\chi_s)\}]^{1/2}$. Other types of parametrization are possible, the main constraint so that (3.3) is a correlation matrix being that $\sum_{s=0}^{S}\varphi^2\{d_r(i,\chi_0)\}=1$ for any i.

Somewhat more generally, we can allow the correlation parameters $(\varrho_1,...,\varrho_S)$ to depend smoothly on covariates, so that $\varrho_s = F\{\Lambda_3(\widetilde{X}_{\chi_s})\}$, where $F(\bullet)$ is a known function, e.g., $F(v)=\exp(v)$. Although very general structures are possible, in our theory we restrict ourselves to the case that $\Lambda_3(\bullet)$ has only one additive term and a scalar smoothing parameter.

B. Penalized Regression Spline Methodology

In this section, we describe the basic fitting methods. Smoothing parameter selection is discussed in Section D. In cases such as ours that computation of a likelihood estimator is infeasible, or at least extremely difficult, it is common to use a composite likelihood formed by adding together individual component loglikelihoods, each of which corresponds to a valid marginal or conditional loglikelihood (Lindsay, 1988). This approach has been used in many problems for correlated binary response data, for example in spatial models (Heagerty and Lele, 1998).

In this section we will discuss penalized composite likelihood estimators for our problem. We assume that each subject has its own mean function but that subjects share the same covariance function. Our asymptotics are as $n_r \to \infty$ for $r=1,...,R$ with R fixed, as is appropriate for our example.

Looking first at individual observations, and defining $\mu^*_{ri} = \mu_{ri}/(1+\sigma^2_\lambda)^{1/2}$, the univariate marginal probability satisfies $\mathrm{pr}(D_{ri}=1|\mu^*_{ri},\sigma^2_\lambda) = \Phi(\mu^*_{ri})$. We define the composite loglikelihood of the first order as a sum of the marginal loglikelihoods pre-

tending that the data are independent. These terms of course give no information about the correlation structure. Thus we use pairs of observations to define a composite likelihood. Let $\Phi_2(\mu_1, \mu_2, \rho)$ be the bivariate standard normal probability of being below μ_1 and μ_2 when the correlation is ρ. Then, pairwise marginal probabilities can be expressed as

$$\text{pr}(D_{ri} = 1, D_{rj} = 1 | \mu^*_{ri}, \mu^*_{rj}, \sigma^2_\lambda, \rho) \;\; = \;\; \Phi_2\{\mu^*_{ri}, \mu^*_{rj}, \sigma^2_\lambda \Omega_{rij}(\rho)/(1 + \sigma^2_\lambda)\}. \quad (3.4)$$

Define $Z^{(kl)}_{rij} = I(D_{ri} = k, D_{rj} = l)$, $\pi^{(kl)}_{rij} = \text{E}(Z^{(kl)}_{rij} | \mu^*_{ri}, \mu^*_{rj}, \sigma^2_\lambda, \rho)$, $\pi^{(1\cdot)}_{rij} = \text{pr}(D_{ri} = 1 | \mu^*_{ri})$ and $\pi^{(\cdot 1)}_{rij} = \text{pr}(D_{rj} = 1 | \mu^*_{rj})$. The composite likelihood of the second order is defined as a sum of loglikelihoods of the second order based on the pairwise distribution of the responses. At least in principle, both the correlation function and the variance of the random effects can be estimated from the pairwise component likelihoods.

Because μ_{ri} and μ^*_{ri} are multiples of one another, in what follows, for notational simplicity, we will simply use the former. We now turn to three general approached based on these ideas.

1. Penalized Composite Likelihood Estimators of the First Order

Recall from (3.2) that the fixed effects for the r^{th} function depend on ζ_{r0} and the two regression splines written as $\Lambda_{rj}(x) = \widetilde{B}^{\text{T}}_{rj}(x)\eta_{rj0}$ for $j = 1, 2$. In order to handle the constraint that $\|\zeta_{r0}\| = 1$ and that its first component is positive, we parameterize to ξ_{r0}, where if ζ_{r0} has $q_{r\zeta}$ components, then ξ_{r0} has $q_{r\zeta} - 1$ components and $\zeta_{r0} = \{(1 - \|\xi_{r0}\|^2)^{1/2}, \xi^{\text{T}}_{r0}\}^{\text{T}}$.

Let β_r be the collection of the parameters $(\eta^{\text{T}}_{r1}, \eta^{\text{T}}_{r2}, \xi^{\text{T}}_r)$, with true value β_{r0}, and write $\mu_{ri} = \mu_{ri}(\beta_r)$. Then pretending that the data are independent, the composite

loglikelihood of first order at the i^{th} location is

$$\mathcal{L}^I_{r,i}(\beta_r) = \log\{\mathbf{L}^I_{r,i}(\beta_r)\} = D_{ri}\log[\Phi\{\mu_{ri}(\beta_r)\}] + (1 - D_{ri})\log[1 - \Phi\{\mu_{ri}(\beta_r)\}].$$

If we sum over the locations for the r^{th} function then we get a composite likelihood function. We need to penalize this function to account for the nonparametric regression, which we do with two smoothing parameters $(\kappa_{r1}, \kappa_{r2})$ and two penalty matrices $(\mathcal{G}_1, \mathcal{G}_2)$. The penalty matrices $(\mathcal{G}_1, \mathcal{G}_2)$ depend on the basis functions used. For example (Ruppert, et al., 2003), for the truncated polynomial series basis of order q and K knots defined in Section 3, $\mathcal{G}_j = \text{diag}(0I_{q_j}, I_{K_j})$, where I_q is the identity matrix of size q. Of course, an interesting subcase is when all functions have the same penalties across rats, i.e., $\kappa_{rj} \equiv \kappa_j$.

For fixed smoothing parameters $(\kappa_{r1}, \kappa_{r2})$, the composite likelihood estimator of the first order is denoted by $\widehat{\beta}^I_r(\kappa_{r1}, \kappa_{r2}) = \widehat{\beta}^I_r$ and is obtained by maximizing

$$\widetilde{\mathcal{L}}^I_r(\beta_r, \kappa_{r1}, \kappa_{r2}) = n_r^{-1}\sum_{i=1}^{n_r}\mathcal{L}^I_{r,i}(\beta_r) - (1/2)\sum_{j=1}^{2}\kappa_{rj}\eta^{\mathrm{T}}_{rj}\mathcal{G}_j\eta_{rj}. \tag{3.5}$$

2. Penalized Composite Likelihood Estimators of the Second Order

In order to estimate the correlation function, we must at least use pairs of observations. We described the models used for the correlation structure in Section 4, which are independent of the function and depend on a bandwidth h and a parameter, η_3, possibly although not necessarily via a spline. Let θ denote the unknown parameters among $(\sigma^2_\lambda, h, \eta_3)$ and let θ_0 denote its true value.

We make the assumption that the covariance structure within a group of animals does not depend on the animal. Let $\mathcal{B}^{\mathrm{T}} = (\beta_1^{\mathrm{T}}, ..., \beta_r^{\mathrm{T}})$ with true value $\mathcal{B}_0$. Define Θ as the collection of all parameters to be estimated, and let Θ_0 be its true value. Organize Θ as $\Theta = (\eta_{11}^{\mathrm{T}}, ..., \eta_{R1}^{\mathrm{T}}, \eta_{12}^{\mathrm{T}}, ..., \eta_{R2}^{\mathrm{T}}, \xi_1^{\mathrm{T}}, ..., \xi_R^{\mathrm{T}}, \theta^{\mathrm{T}})^{\mathrm{T}}$. Using the pairwise probabilities

(3.4), we can write the likelihood at locations (i, j) for function r as

$$\mathcal{L}^{II}_{r,ij}(\Theta) = \sum_{k,\ell=0}^{1} Z^{(k\ell)}_{r,ij} \log\{\pi^{(k\ell)}_{r,ij}(\beta_r, \theta)\}.$$

We propose to maximize a weighted penalized composite likelihood. Let w_{rij} be weights, e.g., the indicator that locations (i, j) are less then specified value apart, a choice that is useful to cut down on the size of the summations and also one that is convenient for later theoretical calculations. Let $\mathcal{W}_r = \sum_{i,j}^{n_r} w_{rij}$ and let $\mathcal{W} = \sum_r \mathcal{W}_r$. Let $\widetilde{\kappa}_j = (\kappa_{1j}, ..., \kappa_{Rj})$ and define $M^{II}(\widetilde{\kappa}_1, \widetilde{\kappa}_2, \kappa_3) = \text{diag}(\kappa_{11}\mathcal{G}_{11}, ..., \kappa_{R1}\mathcal{G}_{R1}, \kappa_{12}\mathcal{G}_{12}, ..., \kappa_{R2}\mathcal{G}_{R2}, 0I_{q^*}, 0I_2, \kappa_3\mathcal{G}^*_3)$, where $q^* = \sum_r q_{r\zeta} - R$ and $\mathcal{G}^*_3$ is diagonal with the penalty matrix for η_3 placed appropriately. For fixed smoothing parameters, the composite likelihood estimator of the second order is denoted by $\widehat{\Theta} = \widehat{\Theta}(\widetilde{\kappa}_1, \widetilde{\kappa}_2, \kappa_3)$ and is obtained by maximizing

$$\widetilde{\mathcal{L}}^{II}(\Theta, \widetilde{\kappa}_1, \widetilde{\kappa}_2, \kappa_3) = \mathcal{W}^{-1} \sum_{r=1}^{R} \sum_{i,j}^{n_r} \{w_{rij}\mathcal{L}^{II}_{r,ij}(\Theta)\} - (1/2)\Theta^{\mathrm{T}} M^{II}(\widetilde{\kappa}_1, \widetilde{\kappa}_2, \kappa_3)\Theta. \quad (3.6)$$

An interesting special case is to have the functions within a group of animals have the same penalty parameters $\kappa_{rk} \equiv \kappa_k$. If the correlation function has no spline component, then $\kappa_3 = 0$.

3. Two-Stage Penalized Estimation of Mean and Association

Maximization of (3.6) can be challenging numerically. This suggests a simple two-stage method. Let $\widehat{\beta}_r$ be the fixed effect parameters obtained from the first order composite likelihood method (3.5) for $r = 1, ..., R$, and let $\widehat{\mathcal{B}}$ be their collection. Define $M^{II*}(\kappa_3) = \text{diag}(0I_2, \kappa_3\mathcal{G}^*_3)$. Then for fixed smoothing parameter κ_3 the two stage estimator $\widehat{\theta}^{II*}(\kappa_3) = \widehat{\theta}^{II*}$ is obtained by maximizing in θ the penalized loglikelihood

$$\widetilde{\mathcal{L}}^{II*}(\theta, \kappa_{r3}) = \mathcal{W}^{-1} \sum_{r=1}^{R} \sum_{i,j}^{n_r} \{w_{rij}\mathcal{L}^{II}_{r,ij}(\widehat{\beta}_r, \theta)\} - (1/2)\theta^T M^{II*}(\kappa_3)\theta. \quad (3.7)$$

C. Asymptotic Results

In this section, we state the main results. All proofs are given in the appendix. Because many of the expressions are lengthy, they too are given in the appendix. We are using an increasing domain asymptotics meaning the the number of sites (grid cells), is increasing asymptotically and the distance between two sites is always greater then some positive number.

We express dependence by means of a model-free mixing coefficient (Guyon, 1995). Define

$$\begin{aligned}\alpha_{u,v}(k) &= \sup\{|\mathrm{pr}(AB) - \mathrm{pr}(A)\mathrm{pr}(B)|, A \in \mathcal{F}(\Lambda_1), B \in \mathcal{F}(\Lambda_2), |\Lambda_1| \leq u, |\Lambda_2| \leq v, \\ &\qquad d(\Lambda_1, \Lambda_2) \geq k\},\end{aligned}$$

where $\Lambda_j \subset \mathcal{Z}^2$, $\mathcal{F}(\Lambda_j)$ is the σ- field generated by $\{D_i, i \in \Lambda_j\}$, $j = 1, 2$, and $d(\Lambda_1, \Lambda_2)$ is the distance between index sets defined as $d(A, B) = \inf\{\max|a_i - b_i| : a = (a_1, a_2) \in A, b = (b_1, b_2) \in B\}$. Define $\alpha(k) = \alpha_{\infty,\infty}(k)$. Throughout, we make the following assumptions.

Assumption 1: For each method, the relevant parameters take values within a compact set, and their true values lie in the interior of that set.

Assumption 2: Define the covariates in the model such as in (3.2) as $\widetilde{X}_{ri}$. The $\widetilde{X}_{ri}$ are assumed to take on values in a compact set, and as $n_r \to \infty$ their empirical distribution function is assumed to converge uniformly to a distribution function.

Assumption 3: For each $r = 1, ..., R$, the random effects $\{\lambda_{r,i}\}_{i=1}^{\infty}$ are α-mixing of size -2 (Gallant, 1987) meaning that the mixing coefficient $\alpha_{\infty,\infty}(k) = o(k^{-2})$.

Assumption 4: The parameters are uniquely identified. Thus for every β_r

(4a) $E\{\widetilde{\mathcal{L}}_r^I(\beta_r, 0, 0)\} \to S_r(\beta_r)$ having a unique maximum at the true value β_{r0},

and for every $\Theta = (\mathcal{B}, \theta)$,

(4b) $E\{\widetilde{\mathcal{L}}^{II}(\Theta, 0, 0, 0)\} \to S(\Theta)$ having a unique maximum at the true value (Θ_0).

Assumption 5: The covariance matrix of the estimating functions are positive definite. Thus,

(5a) For every r, as $n_r \to \infty$, $n_r^{-1} \sum_{i=1}^{n_r} \text{cov}\{\partial \mathcal{L}_{r,i}^{I}(\beta_{r0})/\partial \beta_{r0}\}$,

(5b) As $\min_r n_r \to \infty$, $\mathcal{W}^{-1} \sum_{r=1}^{R} \sum_{i,j}^{n_r} \text{cov}\{w_{rij} \partial \mathcal{L}_{r,ij}^{II}(\Theta_0)/\partial \Theta\}$ are positive definite.

Assumptions 1 and 2 are technical assumptions enabling us to apply certain uniform convergence results. It can be proved that Assumption 3 is satisfied for the Mátern (Stein, 1999) correlation family by exploiting the results of Theorems 1 and 2 of Kolmogorov and Rozanov (1960). Assumption 4 assures us that the parameters are identified. Assumptions 5 is needed to compute asymptotic standard errors.

1. Asymptotic Properties of the First Order Method

We state the result here in some detail because we will need it when we give estimates of the asymptotic covariance matrix. Recall that the fixed effects are given as $\mu_{ri}(\beta_r)$. Define $\mathcal{Z}_{r,i}(\beta_r) = D_{ri} - \Phi\{\mu_{ri}(\beta_r)\}$. Let $\mathcal{Y}_{ri}(\beta_r) = \phi\{\mu_{ri}(\beta_r)\}(\Phi\{\mu_{ri}(\beta_r)\}[1 - \Phi\{\mu_{ri}(\beta_r)\}])^{-1}$. Write $\partial \mu_{ri}(\beta_r)/\partial \beta_r = \mathcal{V}_{ri}(\beta_r)$, and write $v_{r,i}(\beta_r) = \mathcal{V}_{ri}(\beta_r)\mathcal{Y}_{ri}(\beta_r)$. The first derivative of $\mathcal{L}_{r,i}^{I}(\beta_r)$ is $\mathcal{A}_{r,i}^{I}(\beta_r) = v_{r,i}(\beta_r)\mathcal{Z}_{r,i}(\beta_r)$. The derivative of $\mathcal{A}_{r,i}^{I}(\beta_r)$ with respect to β_r and evaluated at β_{r0} is easily seen to have expectation $\mathcal{N}_{ri} = -\phi\{\mu_r(\beta_{r0}, \widetilde{X}_{r,i})\mathcal{Y}_r(\beta_{r0}, \widetilde{X}_{ri})\mathcal{V}_r(\beta_{r0}, \widetilde{X}_{ri})\mathcal{V}_r^{\mathrm{T}}(\beta_{r0}, \widetilde{X}_{ri})$, which is minus the covariance matrix of $\mathcal{A}_{r,i}^{I}(\beta_{r0})$.

Theorem 1: Under Assumptions 1-3, 4(a) and 5(a), if for $j = 1, 2$ the smoothing parameters $\kappa_{rj} = o(1)$, then the first order penalized composite likelihood score equation has a solution $\widehat{\beta}_r^I$ that is a consistent estimator of β_{r0}. In addition, if the smoothing pa-

rameters $\kappa_{rj} = c_{rj} n_r^{-\nu}$, where c_{rj} is finite and $\nu \geq 1/2$, then for the consistent solution of composite likelihood equation, $\widehat{\beta}_r^I$, $(\Sigma_{n,r}^I)^{-1/2} n_r^{1/2} (\widehat{\beta}_r^I - \beta_{r0}) + \Delta_{n,r}^I \beta_{r0} \rightarrow \text{Normal}(0, \text{I})$ in distribution, where $\Delta_{n,r}^I = O(n_r^{-\nu+1/2})$, $\Delta_{n,r}^I$ is defined in the appendix in (B.5) and $\Sigma_{n,r}^I = (\Sigma_{nr,0}^I)^{-1} (\Sigma_{nr,0}^I + \Sigma_{nr,c}^I)(\Sigma_{nr,0}^I)^{-1}$, where $\Sigma_{nr,0}^I$ and $\Sigma_{nr,c}^I$ are defined in the Appendix at (**??**) and (**??**).

2. Asymptotic Properties of the Second Order Method

Theorem 2: Assume that the limit of $\max_r n_r / \min_r n_r$ is finite, as is $\max_r \mathcal{W}_r / \min_r \mathcal{W}_r$. Make assumptions 1-3, 4(b), and 5(b). If all smoothing parameters are of order $o(1)$, then the composite likelihood equation has a solution $\widehat{\Theta}^{II}$ that is a consistent estimator of the true value Θ_0. Assume also that the smoothing parameters $\kappa_{rj} = c_{rj} \mathcal{W}_r^{-\nu}$ and $\kappa_3 = c_3 \mathcal{W}^{-\nu}$, where the c's are finite and $\nu \geq 1/2$. Then for the consistent solution of composite likelihood equation, $\widehat{\Theta}^{II}$, $(\Sigma_{\mathcal{W}}^{II})^{-1/2} \mathcal{W}^{1/2} (\widehat{\Theta}^{II} - \Theta_0) + \Delta_{\mathcal{W}}^{II} \Theta_0 \rightarrow \text{Normal}(0, \text{I})$ in distribution, where $\Delta_{\mathcal{W}}^{II} = O(\mathcal{W}^{-\nu+1/2})$ and the following are defined in the appendix: $\Sigma_{\mathcal{W}}^{II}$ is defined in (B.7) and $\Delta_{\mathcal{W}}^{II}$ is defined in (B.8).

3. Asymptotic Properties of the Two-Stage Method

Theorem 3: Under the assumptions of Theorems 1 and 2, assuming that $\max_r(\mathcal{W}_r / n_r)$ is finite, there is a solution $\widehat{\theta}^{II*}$ that is a strongly consistent estimator of θ_0. In addition, if the smoothing parameters $\kappa_{rj} = c_{rj} n_r^{-\nu}$, $\nu \geq 1/2$ for finite c_{rj}, and if $\kappa_3 = c_3 \mathcal{W}^{-\nu_\theta}$, where c_3 is finite and $\nu_\theta \geq 1/2$, then the consistent solution satisfies $(\Sigma_{\mathcal{W}}^{II*})^{-1/2} \mathcal{W}^{1/2} (\widehat{\theta}^{II*} - \theta_0) + \Delta_{\mathcal{W}}^{II*} \Theta_0 \rightarrow \text{Normal}(0, \text{I})$ in distribution, where $\Delta_{\mathcal{W}}^{II*} = O(\mathcal{W}^{-\nu_\theta+1/2}) + O\{(\min_r n_r)^{-\nu+1/2}\}$ and the following are defined in the appendix: $\Sigma_{\mathcal{W}}^{II*}$ is defined in (B.9) and $\Delta_{\mathcal{W}}^{II*}$ is defined in (B.10).

D. Smoothing Parameter Estimation

It is well-known that with dependent data, standard approaches to smoothing parameter selection, such as cross-validation, lead to undersmoothing, see for example Opsomer, et al. (2001) for discussion and extensive references. The usual device for numerical response data is either to attempt to select the smoothing parameter in such a way as to minimize asymptotic average mean squared error, or more generally to treat the smoothing parameter as a variance component and maximize a resulting mixed model likelihood. Given in our case that the correlated responses are binary and that a likelihood estimate with or without an extra variance component is extremely difficult to compute, at best, it is useful to explore different approaches.

In our problem we estimate parameters by maximizing a composite loglikelihood function rather then minimizing the average squared distance between two curves. Let Θ denote all the parameters, let Θ_0 be their true value and let $\mathrm{CL}(\Theta)$ be a composite loglikelihood without smoothing, as in (3.5)-(3.7). Our approach is to adapt the idea of Kullback-Leibler distance (Kullback and Leibler, 1951) to estimate the smoothing parameters. Define

$$\mathrm{KL}(\Theta, \Theta_0) = \mathrm{E}_{\Theta_0}\{\mathrm{CL}(\Theta_0) - \mathrm{CL}(\Theta)\},$$

where the subscript Θ_0 means that the expectation is taken at the distribution induced by Θ_0. It is technically convenient to work with a symmetrized version of this distance, namely

$$\mathrm{SKL}(\Theta, \Theta_0) = \mathrm{KL}(\Theta, \Theta_0) + \mathrm{KL}(\Theta_0, \Theta).$$

It is easy to see that $\mathrm{SKL}(\Theta, \Theta_0)$ is always non-negative and equals zero when $\Theta = \Theta_0$. If we plug in an estimated $\widehat{\Theta}_\kappa$ into this expression we get a random variable, whose

expectation we then take to find

$$\text{MASKL}(\widehat{\Theta}_\kappa) = \text{E}_{\Theta_0}\{\text{SKL}(\widehat{\Theta}_\kappa, \Theta_0)\}. \tag{3.8}$$

Our goal is to estimate the smoothing parameter so as to minimize (3.8). More precisely, we will find an asymptotically equivalent version of $\widehat{\Theta}_\kappa$, replace it in (3.8) and minimize.

The key to the analysis is that in our problem formulation, $\text{SKL}(\Theta, \Theta_0)$ can be computed analytically. We cannot of course compute (3.8) analytically, but we do show how to compute an asymptotically equivalent version of it, and this allows us to estimate the smoothing parameters. Generally, we must separate the case when the true function is a polynomial of degree $q-1$ or less. We will show that the smoothing parameter associated with such function minimizes the selected criterion when equals to ∞. For other functions, the optimal smoothing parameter is of order $O(n^{-1})$. This is important because it means that there is no asymptotic bias in estimation, i.e., in Theorems 1-3 the terms such as $\Delta^I_{n,r} = 0$.

1. Composite Likelihood of the First Order

For composite likelihood of the first order defined in Section 1, recall that the parameters are $\beta_r = (\eta_{r1}^{\text{T}}, \eta_{r2}^{\text{T}}, \xi_r^{\text{T}},)^{\text{T}}$, with true value β_{r0}. Let the estimate of β_{r0} be $\widehat{\beta}^I_r(\kappa_{r1}, \kappa_{r2})$ With a slight abuse of notation, the fixed effects are $\mu_{ri}(\beta_r)$. It is easily checked that

$$\begin{aligned}\text{SKL}\{\widehat{\beta}^I_r(\kappa_{r1}, \kappa_{r2}), \beta_{r0}\} &= n_r^{-1}\sum_{i=1}^{n_r}(\Phi\{\mu_{ri}(\beta_{r0})\} - \Phi[\mu_{ri}\{\widehat{\beta}^I_r(\kappa_{r1}, \kappa_{r2})\}]) \\ &\quad \times[\varpi_{1,ri}(\beta_{r0}) - \varpi_{1,ri}\{\widehat{\beta}^I_r(\kappa_{r1}, \kappa_{r2})\}],\end{aligned} \tag{3.9}$$

where $\varpi_{2,ri}(\beta) = -\log[1-\Phi\{\mu_{ri}(\beta)\}]$ and $\varpi_{1,ri}(\beta) = \varpi_{2,ri}(\beta) + \log[\Phi\{\mu_{ri}(\beta)\}]$. In the appendix, we sketch an argument indicating that with this definition of SKL($\bullet$), the minimizer of (3.8) when $\eta_{rj0} = 0$, $j = 1, 2$ equals ∞ and in more the general case, the minimizer of (3.8) can be derived as follows. Write $\widetilde{\kappa}^{(r)} = (\kappa_{r1}, \kappa_{r2})^{\mathrm{T}}$. Specifically, let Σ_r be the asymptotic variance of $\widehat{\beta}_r^I$ as in Theorem 1. Then there is a 2×2 matrix $\mathcal{H}_r(\widetilde{\kappa}^{(r)}, \beta_{r0}, \Sigma_{n,r}^I)$ and a 2×1 vector $\mathcal{J}_r(\widetilde{\kappa}^{(r)}, \beta_{r0}, \Sigma_{n,r}^I)$, both of order $O(1)$, such that to terms of order $o(n_r^{-1})$, $\widetilde{\kappa}^{(r)} = n_r^{-1}\mathcal{H}_r^{-1}(\widetilde{\kappa}^{(r)}, \beta_{r0}, \Sigma_{n,r}^I)\mathcal{J}_r(\widetilde{\kappa}^{(r)}, \beta_{r0}, \Sigma_{n,r}^I)$.

Obviously, the question is how to estimate $\Sigma_{n,r}^I$. We take this issue up in detail in Section E. In general, however, estimation of $\Sigma_{n,r}^I$ requires simultaneous estimation of the variance parameters θ_0 through the two-stage method, see Section 3.

2. Composite Likelihood of the Second Order

Let Θ_0 be the true parameter and let $\widehat{\Theta}^{II}(\widetilde{\kappa}_1, \widetilde{\kappa}_2, \kappa_3)$ be the estimator. The, again with a slight abuse of notation, for composite likelihood of the second order, we have that

$$\begin{aligned}\mathrm{SKL}\{\widehat{\Theta}^{II}(\widetilde{\kappa}_1, \widetilde{\kappa}_2, \kappa_3)\} &= \mathcal{W}^{-1}\sum_{r=1}^{R}\sum_{i,j=1}^{n_r}\sum_{k,\ell=0}^{1} w_{rij}[\pi_{r,ij}^{(k\ell)}(\Theta_0) - \pi_{r,ij}^{(k\ell)}\{\widehat{\Theta}^{II}(\widetilde{\kappa}_1, \widetilde{\kappa}_2, \kappa_3)\}] \\ &\quad \times(\log\{\pi_{r,ij}^{(k\ell)}(\Theta_0)\} - \log[\pi_{r,ij}^{(k\ell)}\{\widehat{\Theta}^{II}(\widetilde{\kappa}_1, \widetilde{\kappa}_2, \kappa_3)\}]).\end{aligned}$$

Since in Theorem 2 we have assumed that the sample sizes for each function are proportional, we will write n_0 to be the mean of the sample sizes. In the appendix, we sketch an argument indicating that with this definition of SKL($\bullet$), the minimizer of (3.8) when $\eta_{rj0} \equiv 0$, $j = 1, 2, 3$ equals to ∞ and in more general case, the minimizer of (3.8) is of order $O(n_0^{-1})$. An algorithm similar to that in Section 1 can be defined. Finally, the minimizer of asymptotic MASKL(κ_3) is the solution of the following equation $\widetilde{\kappa}_3 = \mathcal{W}^{-1}\{\mathcal{H}_{\mathcal{W}}^{II*}(\widetilde{\kappa}_3, \mathcal{B}_0, \Sigma_{\mathcal{W}}^{II*})\}^{-1}\mathcal{J}_{\mathcal{W}}^{II*}(\widetilde{\kappa}_3, \mathcal{B}_0, \Sigma_{\mathcal{W}}^{II*})$, where the coefficients are

similar to ones derived before and are given in appendix.

E. Estimation of Asymptotic Covariance Matrices

In this section we will discuss estimation of the asymptotic covariance matrices. Inference and testing are important parts of any data analysis and require an estimation of variance for estimators. Although standard error can be obtained analytically and numerically using multivariate integration, computation is extremely intense and the total number of such calculations can be high. Resampling, bootstrapping and jackknife techniques have been used when direct calculation can be impractical. Theoretical properties of these approaches have been studied for stationary processes, however the form of non-stationarity we have in our case rules out the direct use of mentioned above methods.

In our approach, we will exploit the idea of asymptotic independence: observations that are sufficiently widely separate in space are approximately independent. We will demonstrate the basic idea on standard error estimation for PCL of the first order: the other methods follow with changes of notation. Referring to Theorem 1, it is easy to see that $\Sigma^I_{nr,0}$ can be consistently estimated by $-n_r^{-1}\sum_{i=1}^{n_r}\mathcal{N}_{ri}(\widehat{\beta}_r)$. Next we will discuss the estimation of $\Sigma^I_{nr,0}+\Sigma^I_{nr,c}$. We are going to exploit the fact that the colon is much longer then wide assuming that asymptotically the domain increases only in horizontal direction. Let us define i_x as a horizontal coordinate of site i, $i=1,...,n_r$. Define

$$M\{\beta_r, x_{\max}(n_r)\} = n_r^{-1}[\sum_i \mathcal{A}^I_{r,i}(\beta_r)\{\mathcal{A}^I_{r,i}(\beta_r)\}^T + \sum_{d_x=1}^{x_{\max}(n_r)} q_{d_x,x_{\max}(n_r)} \sum_{i,j:|i_x-j_x|=d_x} \mathcal{A}^I_{r,i}(\beta_r)\times\{\mathcal{A}^I_{r,j}(\beta_r)\}^T],$$

where the weights $q_{d_x,x_{\max}(n_r)}$ are bounded, for any fixed d_x, $\lim_{x_{\max}(\bullet)\to\infty} q_{d_x,x_{max}(\bullet)}=1$

and $x_{max}(n_r) \to \infty$ as $n \to \infty$. The weights are introduced to assure that the estimator is positive definite. For our study we will use Bartlett weights $q_{d,x} = 1 - d/(x+1)$, which is the simplest choice proven to guarantee positivity (Gallant, 1987, page 533). Using Theorem 3 of Gallant (1987, page 534), we have that

$$\begin{aligned} |\mathrm{cov}\{n_r^{-1/2}\sum_j \mathcal{A}^I_{r,j}(\beta_{r,0})\} - \mathrm{E}\{M(\beta_{r,0})\}| &\leq c x^{-1}_{\max}(n_r) \\ \mathrm{pr}(|M(\beta_{r,0}) - \mathrm{E}\{M(\beta_{r,0})\}| > \epsilon) &\leq (c/\epsilon^2) x^4_{\max}(n_r)/n_r. \end{aligned}$$

Therefore for a root-n consistent estimate $\widehat{\beta}_r$, it follows that

$$M(\widehat{\beta}_r, x_{\max}(n_r)) - (\Sigma^I_{nr,0} + \Sigma^I_{nr,c}) = O_p(n_r^{-1/2}) + O_p\{x^4_{\max}(n_r)/n_r\} + O_p\{x^{-1}_{\max}(n_r)\}, \quad c > 0.$$

Note that the optimal rate of convergence for $x_{\max}(n_r)$ is $O(n_r^{1/5})$ in which case $M(\widehat{\beta}_r) - (\Sigma^I_{nr,0} + \Sigma^I_{nr,c}) = O_p(n_r^{-1/5})$. For our numerical work we choose $x_{\max}(n_r) = [n^{1/5}]$, where $[\bullet]$ is an integer part of the real number.

F. Simulation Study

We performed a small simulation study to understand in part the properties of our methods, patterning the study after the analysis of the ACF data that will be presented in Section G. We used the two-stage estimate in this simulation.

In the ACF data, the only covariate is the direction on the grid, X_{ri2}, which we normalize to the unit interval. This means that in (3.2), $\mu_{ri} = \Lambda_{r2}(X_{ri1})$. In addition, instead of the general form (3.3) of the covariance structure, we fit a Mátern correlation model with index 5/2 (Stein, 1999), such that $\Omega_{rij}(\rho) = \exp(-d/\rho)\{1 + d/\rho + (d/\rho)^2/3\}$, where d is the Euclidian distance between sites i and j. Let $\psi = \sigma^2/(1+\sigma^2)$. We run simulations under two different scenarios with correlation parameters chosen

such that $(\psi = 0.5, \Omega(\rho) = 0.5)$ and $(\psi = 0.6, \Omega(\rho) = 0.8)$ thus reflecting moderate dependence of the type found in the ACF data and stronger dependence, respectively. Mimicking the ACF data, we set $n_r = 800$, reflecting a 100×8 grid. The function chosen was $0.5[\sin\{2\pi(X_{ri1} - 0.5)\} - 1]$. For each correlation, we performed 1000 simulations. In this example we used 14-knot cubic splines with truncated polynomial series basis.

There is one smoothing parameter here, κ_{r2}. The algorithm we used was as follows. For each value of κ_{r2}, we solved for the parameters β_{r0} via Fisher scoring. We then updated $\widehat{\theta}$ via Fisher scoring given κ_{r2} and $\widehat{\beta}_r$. Finally, we solved to get an updated κ_{r2}. We iterated until convergence.

Here we study the performance of the algorithm and compare two methods of smoothing parameter selection: the proposed method based on MASKL criterion (Section D) and the traditional one based on cross-validation (CV, Ruppert et.al. 2003). The integrated mean squared errors and squared bias of the probability function for both methods are given in Table 4. Effectively, we see that for estimating the function, both methods are roughly unbiased, but our method has much smaller mean squared errors, with mean squared error efficiencies being roughly 200%. The reason for this is that CV undersmooths the function, leading to increased variability which shows up in individual data sets.

Of course, our algorithm allows us to estimate the correlation along with the mean function, see Table 5.

Included in this table are the mean estimates of $\psi = \sigma^2/(1+\sigma^2)$ and $\Omega(\rho)$, along with the 2.5^{th} and 97.5^{th} percentiles over the simulated data sets.

Table 4.

Results of the simulation. Comparison of estimator performance for different amounts of dependency

$\psi = 0.5$, $\Omega(\rho) = 0.5$		
method	ISB(10^{-2})	IMSE (10^{-2})
MASKL-I	0.003	0.316
MASKL-II	0.008	0.257
CV	0.005	0.553
$\psi = 0.8$, $\Omega(\rho) = 0.6$		
method	ISB(10^{-2})	IMSE (10^{-2})
MASKL-I	0.002	0.689
MASKL-II	0.004	0.490
CV	0.003	1.425

The methods are as follows: 'MASKL' refers to the method based on minimization of MASKL, criterion described in Section D, 'I' and 'II' are one and two-stage algorithms respectively, 'CV' refers to the method based on cross-validation. 'ISB' is integrated squared bias and 'IMSE' is integrated mean squared error.

Table 5.

Results of the simulation. Comparison of proposed algorithm performance when estimating correlation for different amounts of dependency

$\psi = 0.5$, $\Omega(\rho) = 0.5$						
method	$\widehat{\psi}$	$\widehat{\psi}_{.025}$	$\widehat{\psi}_{.975}$	$\widehat{\Omega(\rho)}$	$\widehat{\Omega(\rho)}_{.025}$	$\widehat{\Omega(\rho)}_{.975}$
MASKL-I	0.440	0.190	0.751	0.570	0.252	0.956
MASKL-II	0.507	0.389	0.598	0.441	0.229	0.581
$\psi = 0.8$, $\Omega(\rho) = 0.6$						
method	$\widehat{\psi}$	$\widehat{\psi}_{.025}$	$\widehat{\psi}_{.975}$	$\widehat{\Omega(\rho)}$	$\widehat{\Omega(\rho)}_{.025}$	$\widehat{\Omega(\rho)}_{.975}$
MASKL-I	0.764	0.664	0.972	0.562	0.354	0.858
MASKL-II	0.797	0.728	0.861	0.558	0.465	0.646

The first column is the method: 'MASKL' refers to the method based on minimization of MASKL, criterion described in Section D, while 'I' and 'II' are one and two-stage algorithms respectively. Columns 2-4 and 5-7 are mean, 2.5th, 97.5th percentiles for ψ and $\Omega(\rho)$ respectively

G. Analysis of the ACF Experiment

The introduction mentions the aberrant crypt foci experiment. The details on data collection of the aberrant crypt foci experiment can be found in Apanasovich et. al. (2003). In this study we used $r = 1, ..., 7$ rats in groups of irradiated and non-irradiated animals that were sacrificed at 6 weeks. The only covariate we used was the horizontal distance from the distal part of the colon normalized to the unit interval. In this study we did not use the general random effects structure described in Section D, but fit a Mátern correlation model with index 5/2 (Stein, 1999). We formed the composite likelihood using all pairs of sites less than three units apart, with the choice based on computational feasibility, as well as the fact that even for a relatively high correlation of 0.5 between neatest neighbors, the correlation between sites of more then 3 units apart for the chosen correlation structure is less than 0.02. The rats were allowed their own smoothing parameters. Figure 4 shows the estimated probability of ACF formation as a function of normalized distance from the distal part of the colon for each of the two groups. This figure suggests three important possible conclusions. First, the shapes are complex, and certainly neither constant nor linear. Second, overall the irradiated and non-irradiated groups are different in their ACF formation, with the former having higher ACF formation overall. A simple t-test confirms this finding. Of most potential importance is the finding that ACF formation is not uniform across the colon, and indeed seems greatest at approximately 0.5cm and 0.7cm along the distal colon for the non-irradiated and irradiated rats, respectively. This was initially surprising to us, because we expected that in rats who were allowed to live longer, most tumors would be found in the distal region, rather than roughly in the middle. We went back to other data and confirmed this: there are more tumors roughly where our results suggest most of the ACF occur.

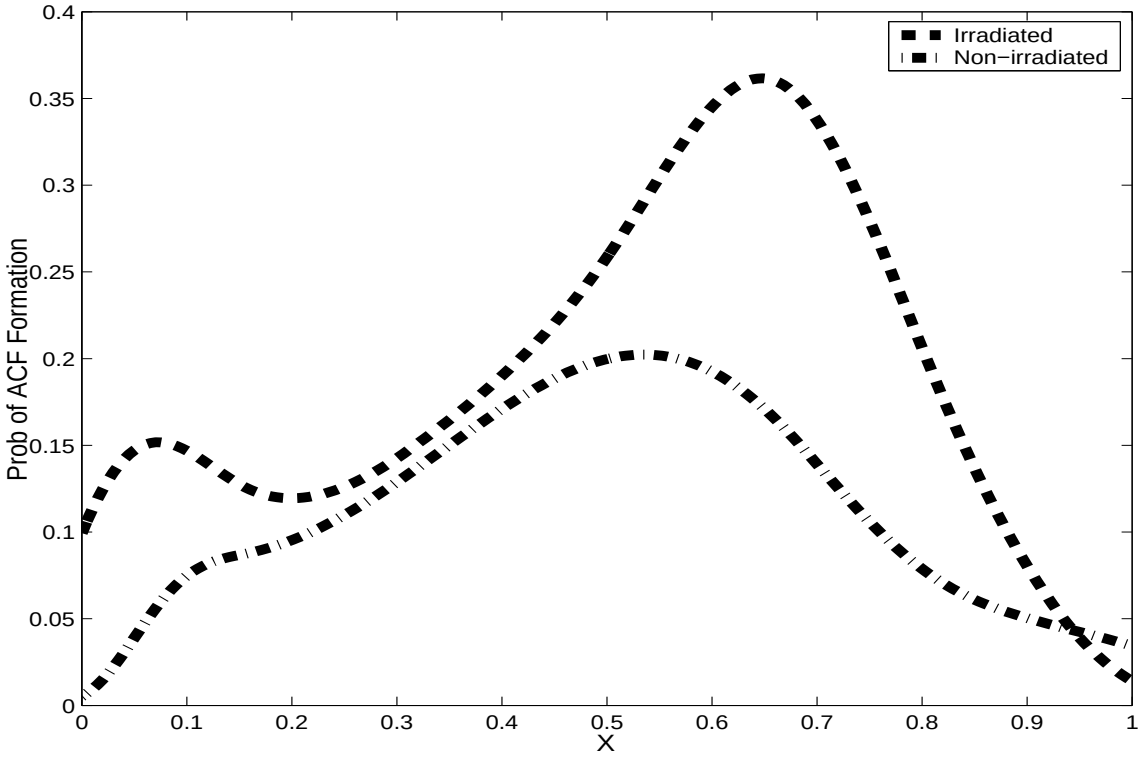

Figure 4. Estimated probabilities of ACF formation. The solid line corresponds to the irradiated group and the dashed line corresponds to the non-irradiated group.

H. Discussion and Extensions

Summarizing briefly, we have described an important experiment in colon carcinogenesis that motivated the study. The responses are binary, they fall into a spatial alignment with marginal probabilities of disease depending strongly on the location within the colon. To handle data like this, we proposed models that have two forms of semiparametrics: (a) one for marginal probabilities, where we studied a combination of partially linear and single index models; and (b) one for the correlation function, where we also proposed a semiparametric model. In all cases, semiparametric modeling of functions was via fixed-knot penalized regression splines with smoothing parameters.

The penalized regression splines have penalty parameters that must converge to zero asymptotically: we derived rates for these parameters that do not lead to an asymptotic bias. We also adapted the idea of Kullback-Leibler distance and derived the optimal rate of convergence for them based on built criteria and proposed a data driven methods to select a proper amount of smoothing. Simulation evidence was positive: even for modeling the mean function our methods did much better than those based on cross-validation.

Finally, we applied the methods to the data from our experiment. We identified regions of high ACF formation that were initially surprising to us but that upon examination of other data actually correspond to regions of high tumor formation. Biologically, this provides a quantification of the localization of ACF formation as precursors to tumors.

CHAPTER IV

CONCLUSION

We have described an important experiment in colon carcinogenesis, where the responses are binary and fall into a spatial alignment, with clear nonstationarity. One key question of interest is whether there is any spatial correlation: its existence would suggest that the response of interest, aberrant crypt foci, are localized in the colon, and thus that regions are affected by radiation and a carcinogen. Our analysis of the aberrant crypt foci experiment suggests that spatial correlation is present at 6–8 weeks after administration of the carcinogen and with or without radiation.

We developed a score–type test for this problem. The original motivation was the Matern class of correlation functions, although we also derived the same test using a particular form of the CAR model. The score–type test method requires no modeling of the correlation per se and is easily computed. We also developed robust score–type tests that bound the influence of a few observations on the score test. The methods are shown via simulation to have test level near the nominal, and also to have in some circumstances more power than a modification of Moran's test.

Assuming that there is correlation, two questions addressed in this study were: the extent of the dependance and the nature of the rate of ACF formation depending on the location within the colon. We proposed a binary mixed model that incorporates a general form of dependency. We assumes that underlying latent process is nonstationary, and we modeled this based on the convolution of latent stationary processes. The dependency of the correlation function on location was also proposed to approximate semiparametrically. We modeled marginal probabilities of ACF formation semiparametrically, using fixed-knot penalized regression splines and single-index models.

We fit the models using pairwise pseudolikelihood methods. Assuming that the underlying latent process is strongly mixing, known to be the case for many Gaussian processes, we proved asymptotic normality of the methods. The penalized regression splines have penalty parameters that must converge to zero asymptotically: we derived rates for these parameters that do not lead to an asymptotic bias. We also adapted the idea of Kullback-Leibler distance and derived the optimal rate of convergence for them based on built criteria and proposed a data driven methods to select a proper amount of smoothing. The method is shown via simulations to produce unbiased estimators. It was demonstrated that they are less variable than estimators obtained by using cross-validation for the smoothing parameter selection.

Finally, we applied the methods to the data from our experiment. We identified regions of high ACF formation that actually correspond to regions of high tumor formation. This provides some more evidence that ACF are precursor lesions of experimental colon cancer. From a practical point of view, ACF are in fact early biomarkers of cancer risk and should be used in innervation studies aimed at identifying agents able to reduce the incidence of colorectal carcinoma. Thus it is useful to look for ACF in colons in order to shed some light on the earliest evens of colon carcinogenesis and to test measures to prevent colorectal cancer.

REFERENCES

Abramowitz, M. and Stegun, I. (1965). *Handbook of Mathematical Functions,*Dover, New York.

Andrews, D. W. K. (1991). An empirical process central limit theorem for dependent non-identically distributed random variables. *Journal of Multivariate Analysis*, **38**, 187-203.

Apanasovich, T. V., Sheather, S., Lupton, J. R., Popovic, N., Turner, N. D., Chapkin, R. S. and Carroll, R. J. (2003). Testing for spatial correlation in nonstationary binary data with application to aberrant crypt foci in colon carcinogenesis. *Biometrics*, **59**, 752-761.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B,* **36**, 192–236.

Bird, R. P. (1995). Role of aberrant crypt foci in understanding the pathogenesis of colon cancer. *Cancer Letters*,**93**, 55-71.

Bird, R. P. and Good, C. K. (2000). The significance of aberrant crypt foci in understanding the pathogenesis of colon cancer. *Toxicology Letters*, **112-113**, 395-402.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*,**88**, 9–36.

Carroll, R. J. and Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society, Series B*, **55**, 693–706.

Cliff, A. D. and Ord, J. K. (1981). *Spatial Processes: Models and Applications.* Pion Limited, London.

Commenges, D. and Jacqmin-Gadda, H. (1997). Generalized score test of homogene-

ity based on correlated random effects models. *Journal of the Royal Statistical Society, Series B*,**59**, 157–171.

Diggle, P. J., Moyeed, R. A. and Tawn, J. A. (1997). Model-based geostatistics. *Applied Statistics*, **47**, Part 3, 299–350.

Eilers, P. H. C. and Marx,B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **11**, 89-121.

Fuentes, M., (2002). Spectral methods for nonstationary spatial processes *Biometrika*, **89**, 197-210.

Jacqmin-Gadda, H., Commenges, D., Nejjari, C. and Dartigues, J. F. (1997). Tests of geographical correlation with adjustment for explanatory variables: an application to dyspnoea in the elderly. *Statistics in Medicine*, **16**, 1283–1297.

Gallant, R. A. (1987). *Nonlinear Statistical Models.* Wiley, New York.

Guyon, X. (1995). *Random Fields on a Network: Modelling, Statistics, and Applications*. Springer-Verlag Inc, Berlin.

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to biniry spatial data. *Journal of the American Statistical Association*, **93**, 1099-1123.

Kolmogorov, A. N. and Rozanov, Yu. A. (1960). On strong mixing conditions for stationary Gaussian processes. *Theory of Probability and Its Applications*, **5**, 204-207.

le Cessie, S. and van Houwelingen, J. C. (1994). Logistic regression for correlated binary data. *Applied Statistics*, 43, 95–108.

Lindsay, B. G. (1988). Composite likelihood methods. *Statistical Inference for Stochastic Processes*, **80**, 221-239.

McLellan, E. A., Medline, A. and Bird, R. P. (1991). Sequential analyses of the growth and morphological characteristics of aberrant crypt foci: putative preneoplastic lesions. *Cancer Research*, **51**, 5270–5274.

Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B*, **10**, 54–62.

Morris, J. S., Wang, N., Lupton, J. R., Chapkin, R. S., Turner, N. D. Hong, M. Y. & Carroll, R. J. (2001). Parametric and nonparametric methods for understanding the relationship between carcinogen–induced DNA adduct levels in distal and proximal regions of the colon. *Journal of the American Statistical Association*, **96**, 816–826.

Morris, J. S., Wang, N., Lupton, J. R., Chapkin, R. S., Turner, N. D. Hong, M. Y. & Carroll, R. J. (2002). A Bayesian analysis of colonic crypt structure and coordinated response incorporating missing crypts. *Biostatistics*, **3**, 529–546.

Morris, J. S., Vannucci, M., Brown, P. J. & Carroll, R. J. (2003). Wavelet–based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association*, to appear.

Opsomer, J., Wang, Y. and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, **16**, 134-153.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* Wiley, New York.

Richardson S., Guihenneuc C. and Lasserre V. (1992). Spatial linear models with autocorrelated error structure. *Statistician*, **41**, 539–557.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, **11**, 735-757.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge University Press, Cambridge.

Simpson, D. G., Ruppert, D. and Carroll, R. J. (1992). One–step GM–estimates and stability of inferences in linear regression. *Journal of the American Statistical Association*, **87**, 439–450.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* Springer-Verlag Inc, Berlin.

White, H. (1984). *Asymptotic Theory for Econometricians.* Academic Press, Orlando.

Withers, C. S. (1981). Central limit theorems for dependent variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **57**, 509-534.

Wooldridge, J. (1986). *Asymptotic Properties of Econometric Estimators.* Ph.D. dissertation, Department of Economics, University of California, San Diego.

APPENDIX A

Recall that there are $k = 1, ..., N$ pairs of responses (Y_{1k}, Y_{2k}), and that $Z_{ijk} = I(Y_{1k} = i, Y_{2k} = j)$. Note that $\mathrm{E}(Z_{ijk}) = \pi_{ijk}$. Also, $\pi_{01k} = \pi_{\cdot 1k} - \pi_{11k}$, $\pi_{10k} = \pi_{1\cdot k} - \pi_{11k}$, $\pi_{00k} = 1 - \pi_{01k} - \pi_{10k} - \pi_{11k} = 1 - \pi_{\cdot 1k} - \pi_{1\cdot k} + \pi_{11k}$. In addition, it follows that $\partial\pi_{00k}/\partial\rho = \partial\pi_{11k}/\partial\rho$, $\partial\pi_{01k}/\partial\rho = -\partial\pi_{11k}/\partial\rho$ and $\partial\pi_{10k}/\partial\rho = -\partial\pi_{11k}/\partial\rho$. Then the differentiation of the loglikelihood with respect to ρ leads to

$$\frac{\partial \mathrm{logL}(\rho)}{\partial\rho} = \sum_k \left(\frac{Z_{00k}}{\pi_{00k}} - \frac{Z_{01k}}{\pi_{01k}} - \frac{Z_{10k}}{\pi_{10k}} + \frac{Z_{11k}}{\pi_{11k}} \right) \frac{\partial\pi_{11k}}{\partial\rho}.$$

The information can be shown to equal

$$\mathrm{E}\left\{ -\frac{\partial^2 \mathrm{logL}(\rho)}{\partial\rho^2} \right\} = \sum_k \left(\frac{1}{\pi_{00k}} + \frac{1}{\pi_{01k}} + \frac{1}{\pi_{10k}} + \frac{1}{\pi_{11k}} \right) \left(\frac{\partial\pi_{11k}}{\partial\rho} \right)^2.$$

Recall that $\psi = \sigma_\lambda^2/(1 + \sigma_\lambda^2)\Omega$. If $\pi_{11k} = \Phi_2\{\mu_{1k}^*, \mu_{2k}^*, \psi_k(\rho)\}$, then

$$\frac{\partial\pi_{11k}}{\partial\rho} = \frac{\partial\pi_{11k}}{\partial\psi_k}\frac{\partial\psi_k}{\partial\rho} = \phi_2\{\mu_{1k}^*, \mu_{2k}^*, \psi_k(\rho)\}\frac{\partial\psi_k}{\partial\rho},$$

where $\phi_2\{\mu_{1k}^*, \mu_{2k}^*, \psi_k(\rho)\}$ is a bivariate standard normal density function with correlation $\psi_k(\rho)$, evaluated at (μ_{1k}^*, μ_{2k}^*). Therefore the score is

$$\left(\frac{Z_{00k}}{\pi_{00k}} - \frac{Z_{01k}}{\pi_{01k}} - \frac{Z_{10k}}{\pi_{10k}} + \frac{Z_{11k}}{\pi_{11k}} \right) \phi_2\{\mu_{1k}^*, \mu_{2k}^*, \psi_k(\rho)\}\frac{\partial\psi_k}{\partial\rho}$$

When we use as the correlation function, $\Omega_k(\rho)$, a member of the Matern class with $\nu = m + \frac{1}{2}$, the correlation function is of the form $\exp(-d_k/\rho)$ times a polynomial in $-d_k/\rho$ degree m (Abramowitz and Stegun 1965, 10.2.15). The derivative of such correlation function is of the form $\exp(-d_k/\rho)$ times a polynomial in $-d_k/\rho$ degree $m + 2$. Using L'Hospital's rule one can prove that $\exp(-d_k/\rho)(-d_k/\rho)^l \to 0$ as

$\rho \to 0$ for $l \geq 0$. Hence $\partial\psi_k(\rho)/\partial\rho = \sigma_\lambda^2/(1+\sigma_\lambda^2)\partial\Omega_k(\rho)/\partial\rho \to 0$ as $\rho \to 0$ and as the result the score evaluated at $\rho = 0$ goes to 0 as well. The solution to that technical difficulty we propose is to focus on pairs that are exactly the same distance apart, in which case $\psi_k(\rho)$ is going to be the same for all k. We reparameterize $\psi_1(\rho) = \psi_2(\rho) = ... = \psi_N(\rho) = \psi$. Notice that $\psi = 0$ if and only if $\rho = 0$. So we can reformulate the null hypothesis of no spatial correlation in terms of a new parameter, $H_0 : \psi = 0$. The new score that will be used in constructing the test is also recalculated in terms of a new parameter ψ

$$s_k(\mu_{1k}^*, \mu_{2k}^*, \psi) = \left(\frac{Z_{00k}}{\pi_{00k}} - \frac{Z_{01k}}{\pi_{01k}} - \frac{Z_{10k}}{\pi_{10k}} + \frac{Z_{11k}}{\pi_{11k}}\right) \phi_2\{\mu_{1k}^*, \mu_{2k}^*, \psi\} \tag{A.1}$$

This shows that the reparametrization allows us to eliminate the term $\partial\psi_k/\partial\rho$ which is equal to 0 when $\rho = 0$. Notice that $\pi_{ijk}|_{\psi=0} = (-1)^{i+j}(1-i-\pi_{1\cdot k})(1-j-\pi_{\cdot 1k})$ and $\phi_2\{\mu_{1k}^*, \mu_{2k}^*, 0\} = \phi(\mu_{1k}^*)\phi(\mu_{1k}^*)$. Hence the score evaluated at $\psi = 0$

$$\mathcal{G}_k(\mu_{1k}^*, \mu_{2k}^*) = s_k(\mu_{1k}^*, \mu_{2k}^*, 0) = \frac{(Y_{1k} - \pi_{1\cdot k})(Y_{2k} - \pi_{\cdot 1k})\phi(\mu_{1k}^*)\phi(\mu_{2k}^*)}{\pi_{1\cdot k}(1-\pi_{1\cdot k})\pi_{\cdot 1k}(1-\pi_{\cdot 1k})}.$$

The variance of $\mathcal{G}_k(\mu_{1k}^*, \mu_{2k}^*)$ under the null hypothesis is

$$\text{var}\{\mathcal{G}_k(\mu_{1k}^*, \mu_{2k}^*)\} = \frac{\{\phi(\mu_{1k}^*)\phi(\mu_{2k}^*)\}^2}{\pi_{1\cdot k}(1-\pi_{1\cdot k})\pi_{\cdot 1k}(1-\pi_{\cdot 1k})}.$$

Let the μs depend on covariates X and a parameter β_*, i.e., $\mu(X, \beta_*)$ with the property that for any constant c, $c\mu(X, \beta_*) = \mu(X, \beta_{**})$ for some β_{**}. Recall that under the null hypothesis the Ys are independent and $\text{pr}(Y = 1|X) = \Phi\{\mu(X, \beta_*)/(1+\sigma_\lambda^2)^{1/2}\} = \Phi\{\mu(X, \beta)\}$. Then the score can be written as

$$\mathcal{H}_k(\beta) = \frac{[Y_{1k} - \Phi\{\mu(X_{1k}, \beta)\}][Y_{2k} - \Phi\{\mu(X_{2k}, \beta)\}]\phi\{\mu(X_{1k}, \beta)\}\phi\{\mu(X_{2k}, \beta)\}}{\Phi\{\mu(X_{1k}, \beta)\}[1 - \Phi\{\mu(X_{1k}, \beta)\}]\Phi\{\mu(X_{2k}, \beta)\}[1 - \Phi\{\mu(X_{2k}, \beta)\}]}.$$

The variance of $\mathcal{H}_k(\beta)$ under the hypothesis of no spatial correlation is

$$\mathcal{V}_k(\beta) = \frac{[\phi\{\mu(X_{1k},\beta)\}\phi\{\mu(X_{2k},\beta)\}]^2}{\Phi\{\mu(X_{1k},\beta)\}[1-\Phi\{\mu(X_{1k},\beta)\}]\Phi\{\mu(X_{2k},\beta)\}[1-\Phi\{\mu(X_{2k},\beta)\}]}.$$

Then our test statistic is

$$\frac{\sum_k \mathcal{H}_k(\beta)}{\{\sum_k \mathcal{V}_k(\beta)\}^{1/2}}. \tag{A.2}$$

There is an important subtlety in (A.2), namely that while not independent, under the null hypothesis the terms $\mathcal{H}_k(\beta)$ are mutually uncorrelated, and hence (A.2) indeed has mean 0 and variance 1.

That (A.2) is asymptotically normally distributed with mean zero and variance one under the null hypothesis of independence follows from a result of Commenges and Jacqmin–Gadda (1997), under conditions that govern the behavior of the terms $\mu(X,\beta)$.

Of course, β is not known. The result of Commenges and Jacqmin–Gadda (1997) can be used to show that when a $\sqrt{n}$–consistent estimate $\widehat{\beta}$ is substituted into the test statistic in place of the true value β, the limit distribution of the test statistic is unaffected. Because of page limits we do not provide details, but the essential point is the orthogonality of the numerator of the test to β, i.e., it may be shown that under the null hypothesis of independence,

$$\mathrm{E}\left\{\frac{\partial \mathcal{H}_k(\beta)}{\partial \beta}\right\}|_{\psi=0} = 0.$$

Therefore our test statistic

$$\frac{\sum_k \mathcal{H}_k(\widehat{\beta})}{\{\sum_k \mathcal{V}_k(\widehat{\beta})\}^{1/2}}$$

is asymptotically standard normal under the null hypothesis of independence. We now show that our score test is also the score test for the conditional autoregressive model

(CAR), see Besag (1974) and Richardson, et al. (1992). Let $\widetilde{\lambda}$ be the vector of the λs. For the Gaussian CAR model, $\lambda = \text{Normal}\{0, \sigma_\lambda^2 B(\rho)\}$, where $B(\rho) = (I - \rho C)^{-1}$, where C is chosen to be a neighborhood matrix whose (i,j)th element is equal to 1 if region i and region j $(i \neq j)$ are neighbors. Let assume that for the set of locations $(1, ..., n)$ we observe a binary vector $(Y_1, Y_2, ..., Y_n)$. Define the likelihood function

$$\mathrm{L}(\rho) = \mathrm{pr}(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n) = \int ... \int \phi\{x_1, x_2, ..., x_n; I + \sigma_\lambda^2 B(\rho)\} dx_1, dx_2, ..., dx_n,$$

where $\phi(x_1, x_2, ..., x_n; V)$ is the n–variate normal density with mean 0 and covariance matrix V, the integral with respect to x_i is from $-\infty$ to μ_i if $Y_i = 1$ and from μ_i to ∞ if $Y_i = 0$.

Let the elements of $B(\rho)$ be $B_{ij}(\rho)$. Make the change of variables $z_i = x_i / \{1 + \sigma_\lambda^2 B_{ii}(\rho)\}^{1/2}$, so that $\mathrm{L}(\rho) = \mathrm{pr}(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n) = \int ... \int \phi\{z_1, z_2, ..., z_n; \Sigma(\rho)\}\, dz_1, dz_2, ..., dz_n$, where Σ has elements Σ_{ij}, with $\Sigma_{ii} = 1$ and $\Sigma_{ij}(\rho) = \sigma_\lambda^2 B_{ij}(\rho) / [\{1 + \sigma_\lambda^2 B_{ii}(\rho)\}\{1 + \sigma_\lambda^2 B_{jj}(\rho)\}]^{1/2}$, and the integral with respect to z_i is from $-\infty$ to μ_i^* if $Y_i = 1$ and from μ_i^* to ∞ if $Y_i = 0$, where $\mu_i^* = \mu_i / \{1 + \sigma_\lambda^2 B_{ii}(\rho)\}^{1/2}$. The first derivative of the loglikelihood with respect to ρ is

$$\begin{aligned}
\frac{\partial \log \mathrm{L}(\rho)}{\partial \rho} &= \frac{1}{\mathrm{L}(\rho)} \frac{\partial \mathrm{L}(\rho)}{\partial \rho} = \frac{1}{\mathrm{L}(\rho)} \left(\sum_{i<j} \frac{\partial \mathrm{L}(\rho)}{\partial \Sigma_{ij}} \frac{\partial \Sigma_{ij}}{\partial \rho} + \sum_i \frac{\partial L(\rho)}{\partial \mu_i^*} \frac{\partial \mu_i^*}{\partial \rho} \right) \\
&= \frac{1}{\mathrm{L}(\rho)} \sum_{i<j} (2Y_i - 1)(2Y_j - 1) \int ... \int \phi\{z_1, ..., \mu_i^*, ..., \mu_j^*, ..., z_n; \Sigma(\rho)\} \\
&\quad \times \prod_{k \neq i,j} dz_k \frac{\partial \Sigma_{ij}}{\partial \rho} - \frac{1}{\mathrm{L}(\rho)} \sum_i \frac{\partial L(\rho)}{\partial \mu_i^*} \frac{\mu_i}{2\{1 + \sigma_\lambda^2 B_{ii}(\rho)\}^{3/2}} \sigma_\lambda^2 \left(\frac{\partial B}{\partial \rho} \right)_{ii}.
\end{aligned}$$

Now note that when $\rho = 0$, $B = I$. Also, $\partial B(\rho)/\partial \rho|_{\rho=0} = C$, with diagonal elements equal to 0. This means that

$$\frac{\partial \Sigma_{ij}}{\partial \rho}|_{\rho=0} = \frac{\sigma_\lambda^2 C_{ij}}{1 + \sigma_\lambda^2}.$$

Now note that $\Sigma(\rho = 0) = I$, the identity matrix. Using the previous results, it follows that

$$\begin{aligned}\frac{1}{\mathrm{L}(\rho)}\frac{\partial \mathrm{L}(\rho)}{\partial \Sigma_{ij}}|_{\rho=0} &= \frac{(2Y_i-1)(2Y_j-1)\phi(\mu_i^*)\phi(\mu_j^*)\prod_{k\neq i,j}\mathrm{pr}(Y_k=y_k)}{\prod_k \mathrm{pr}(Y_k=y_k)} \\ &= \frac{(2Y_i-1)(2Y_j-1)\phi(\mu_i^*)\phi(\mu_j^*)}{\mathrm{pr}(Y_i=y_i)\mathrm{pr}(Y_j=y_j)} = \frac{(Y_i-\pi_i)(Y_i-\pi_i)\phi(\mu_i^*)\phi(\mu_j^*)}{\pi_i(1-\pi_i)\pi_j(1-\pi_j)}.\end{aligned}$$

Hence

$$\frac{\partial \mathrm{logL}(\rho)}{\partial \rho}|_{\rho=0} = \sum_{i<j:C_{ij}\neq 0}\frac{(Y_i-\pi_i)(Y_i-\pi_i)\phi(\mu_i^*)\phi(\mu_j^*)}{\pi_i(1-\pi_i)\pi_j(1-\pi_j)}\frac{\sigma_\lambda^2}{(1+\sigma_\lambda^2)}.$$

The common term $\sigma_\lambda^2/(1+\sigma_\lambda^2)$ can be omitted from the Score Test statistic, leading to our test, as claimed.

APPENDIX B

Definition: Two norms that are going to be used are defined as follows: if $X = \{X_i\}_{i=1}^k$ is a k-vector valued random variable, mapping a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ into $\mathcal{R}^k$, then

$$\begin{aligned} ||X||_r &= \left\{ \sum_{i=1}^k \int_\Omega |X_i(\omega)|^r dP(\omega) \right\}^{1/r}; \\ ||X|| &= \left(\sum_{i=1}^k X_i^2 \right)^{1/2}. \end{aligned}$$

Definition: For two random sequencies X_n and Y_n, $X_n = o_p(Y_n)$ means X_n/Y_n converges to 0 in probability, $X_n = O_p(Y_n)$ means the sequence $\{X_n/Y_n\}_{n=1}^\infty$ is tight, for any $\epsilon > 0$ there exists a constant $M > 0$ such that $\mathrm{pr}(|X_n/Y_n| \leq M) \geq 1 - \epsilon$.

All definitions below are from Gallant (1987).

Definition: A sequence $\{\alpha_m\}_{m=1}^\infty$ of nonnegative real numbers is said to be of **size** $-q$ if $\alpha_m = O(m^\Theta)$ for some $\Theta < -q$

Definition: A measure of dependence between two σ-algebras $\mathcal{F}$ and $\mathcal{G}$ is

$$\alpha(\mathcal{F}, \mathcal{G}) = \sup_{F \in \mathcal{F}, G \in \mathcal{G}} |\mathrm{pr}(FG) - \mathrm{pr}(F)\mathrm{pr}(G)|.$$

Let $\{V_t\}_{t=-\infty}^\infty$ be a sequence of random variables defined on the complete probability space $(\Omega, \mathcal{A}, \mathrm{P})$ described above, and let

$$\mathcal{F}_m^n = \sigma(V_m, V_{m+1}, ..., V_n)$$

denote the smallest complete (with respect to P) sub-σ-algebra such t hat the random

variables V_t for $t = m, m+1, ..., n$ are measurable. Define

$$\alpha_m = \sup_t \alpha(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^\infty).$$

$\{V_t\}_{t=-\infty}^\infty$ is **strong mixing** or **α-mixing** if $\alpha_m \to 0$ as $m \to \infty$.

Definition: Another measure of dependence between two σ-algebras $\mathcal{F}$ and $\mathcal{G}$ is

$$\rho(\mathcal{F}, \mathcal{G}) = \sup_{F \in \mathcal{F}, G \in \mathcal{G}} |\text{correlation}(F, G)|.$$

Let $\{V_t\}_{t=-\infty}^\infty$ be a sequence of random variables defined on the complete probability space $(\Omega, \mathcal{A}, \mathrm{P})$ described above, and let

$$\mathcal{F}_m^n = \sigma(V_m, V_{m+1}, ..., V_n)$$

denote the smallest complete (with respect to P) sub-σ-algebra such that the random variables V_t for $t = m, m+1, ..., n$ are measurable. Define

$$\rho_m = \sup_t \rho(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^\infty).$$

$\{V_t\}_{t=-\infty}^\infty$ is **ρ - mixing** if $\rho_m \to 0$ as $m \to \infty$.

Definition: Let $\{V_t\}_{t=-\infty}^\infty$ be a sequence of vector valued random variables defined on the complete probability space $(\Omega, \mathcal{A}, P)$, and let $\mathcal{F}_m^n$ denote the smallest complete sub-σ-algebra such that the random variables V_t for $t = m, m+1, ..., n$ are measurable. Let $W_t = W_t(V_\infty)$ for $t = 0, 1, ...$ denote a sequence of Borel measurable functions with range in $\mathcal{R}^k$ that depend (possibly) on infinitely many of the coordinates of the vector $V_\infty = (..., V_{-1}, V_0, V_1, ...)$. Let $\{g_{nt}(w_t)\}$ for $n = 1, 2, ...$ and $t = 0, 1, 2, ...$ be a doubly indexed sequence of real valued, Borel measurable functions each of which is defined over $\mathcal{R}^k$. The doubly indexed sequence $\{g_{nr}(W_t)\}$ is said to be **near epoch**

dependent of size $-q$ if

$$v_m = \sup_n \sup_t ||g_{nr}(W_t) - \mathrm{E}\{g_{nt}(W_t)|\mathcal{F}_{t-m}^{t+m}\}||_2$$

is of size $-q$.

Note that if W_t depends on only a finite number of the V_t, $W_t = W_t(V_{t\pm M})$, $M < \infty$ then any sequence $g_{nt}(w_t)$ will be near epoch dependent, because

$$||g_{nt}(W_t) - \mathrm{E}\{g_{nt}(W_t)|\mathcal{F}_{t-m}^{t+m}\}||_2 = 0.$$

Definition: Let $\{W_t\}_{t=0}^{\infty}$ be a sequence of random variables defined on the probability space $(\Omega, \mathcal{A}, \mathrm{P})$, each with range in $\mathcal{R}^{k_t}$. A sequence of functions $\{g_t(W_t, \gamma)\}$ defined over a metric space (Γ, ρ) is **$\mathcal{A}$-smooth** if for each γ in Γ there is a constant $\delta > 0$ such that $||\gamma - \gamma_0|| \leq \delta$ implies

$$|g_t(W_t, \gamma) - g_t(W_t, \gamma_0)| \leq B_t(W_t)h\{\rho(\gamma, \gamma_0)\}$$

except on some set $E \in \Omega$ with $\mathrm{pr}(E) = 0$ where $B_t : \mathcal{R}^{k_t} \to \mathcal{R}^+$ and $h : \mathcal{R}^+ \to \mathcal{R}^+$ are nonrandom functions such that $B_t(w_t)$ is Borel measurable,

$$\frac{1}{n}\sum_{t=1}^{n} \mathrm{E}\{B_t(W_t)\} \leq \Delta < \infty \text{ for all } n$$

and $h(x) \to h(0) = 0$ as $x \to 0$; δ, $B_t(\bullet)$ and may $h(\bullet)$ depend on γ_0.

Definition: Let $\{X_{nt} : n = 1, 2, ...; t = 1, 2, ...\}$ be a doubly indexed sequence of real valued random variables in $L_2(\Omega, \mathcal{A}, \mathrm{P})$, and let $\mathcal{F}_{-\infty}^{t}$ be an increasing sequence of sub-σ-algebras. Then $(X_{nt}, \mathcal{F}_{-\infty}^{t})$ is a **mixingale** if for sequences of nonnegative constants $\{c_{nt}\}$ and $\{\psi_m\}$ with $\lim_{m\to\infty} \psi_m = 0$ we have for all $t \geq 1$, $n \geq 1$, and $m \geq 0$ that

$$||\mathrm{E}(X_{nt}|\mathcal{F}_{-\infty}^{t-m})||_2 \leq \psi_m c_{nt};$$

$$||X_{nt} - \mathrm{E}(X_{nt}|\mathcal{F}_{-\infty}^{t+m})||_2 \leq \psi_{m+1} c_{nt}.$$

In what follows, we will need a central limit theorem. We use a result due to Wooldridge (1986), cited by Andrews (1991, Proposition 1) and similar to an approach of Withers (1981).

Central Limit Theorem Let $V_\infty = \{V\}_{i=-\infty}^{\infty}$ be a sequence of vector-valued random variables that are strong mixing of size $-2q/(q-2)$ for some $q > 2$. Let $W_i = W_i(V_\infty)$ denote a sequence of functions with range in $\mathcal{R}^{k_i}$, that depends on finitely many of the coordinates of the V's. Let $\{g_{ni}(W_i)\}_{i=1}^{\infty}$ be a sequence of real-valued functions with $\mathrm{var}\{n^{-1/2}\sum_{i=1}^{n} g_{ni}(W_i)\} = \sigma_n^2$ and $\sup_i \mathrm{E}|g_{ni}(W_i)|^r < \infty$

(a) if $\lim_{n\to\infty} \sigma_n^2 = \sigma^2 < \infty$, then $(n\sigma^2)^{-1/2}\sum_{i=1}^{n}[g_{ni}(W_i) - \mathrm{E}\{g_{ni}(W_i)\}]$ converges in distribution to Normal$(0,1)$.

(b) if $\liminf_{n\to\infty} \lambda_{\min}(\sigma_n^2) > 0$, where $\lambda_{\min}$ is the smallest eigenvalue then $(n\sigma_n^2)^{-1/2}\sum_{i=1}^{n}\ [g_{ni}(W_i) - \mathrm{E}\{g_{ni}(W_i)\}]$ converges in distribution to Normal$(0,1)$.

Refer to the statement of Theorem 1 for definitions. From Assumption 2, let the limit of the empirical distribution function of the covariates $\widetilde{X}_{ri}$ be $F_{Gr}(\widetilde{x})$. Define $M(\kappa_{r1}, \kappa_{r2}) = \mathrm{diag}(\kappa_{r1}\mathcal{G}_{r1}, \kappa_{r2}\mathcal{G}_{r2}, 0I_{q_\zeta - 1})$. The first derivative of $\widetilde{\mathcal{L}}_{r,i}^{I}(\beta_r)$ is

$$n_r^{-1}\sum_{i=1}^{n_r} \mathcal{A}_{r,i}^{I}(\beta_r) - M(\kappa_{r1}, \kappa_{r2})\beta_r. \tag{B.1}$$

Proof of Theorem 1:

Using the Uniform Law of Large Numbers (Gallant, 1987, page 515, Theorem 1), and using Assumptions 1 and 2 and the fact that the response are binary and hence bounded, it is readily verified that $\lim_{n_r\to\infty} \sup_{\beta_r\in\mathcal{B}} |n_r^{-1}\sum_j \mathcal{L}_{r,i}^{I}(\beta_r) - n_r^{-1}\sum_j \mathrm{E}\{\mathcal{L}_{r,i}^{I}(\beta_r)\}| = 0$ almost surely. Moreover, once again remembering the com-

pactness assumptions 1 and 2, it then follows that

$$\lim_{n_r\to\infty}\sup_{\beta_r\in\mathcal{B}}|n_r^{-1}\sum_{i=1}^{n_r}(\mathcal{L}^I_{r,i}(\beta_r)-\mathcal{L}^I_{r,i}(\beta_{r0})-[\mathrm{E}\{\mathcal{L}^I_{r,i}(\beta_r)\}-\mathrm{E}\{\mathcal{L}^I_{r,i}(\beta_{r0})\}])|=0$$

almost surely. It is clear that $\mathrm{E}[\mathbf{L}^I_{r,i}(\beta_r)/\mathbf{L}^I_{r,i}(\beta_{r0})]=1$. Since logarithm is a strictly concave function, therefore by Jensen's inequality

$$\begin{aligned}\sup_{\beta_r}[\mathrm{E}\{\mathcal{L}^I_{r,i}(\beta_r)\}-\mathrm{E}\{\mathcal{L}^I_{r,i}(\beta_{r0})\}] &= \sup_{\beta_r}\mathrm{E}[\log\{\mathbf{L}_{r,i}(\beta_r)/\mathbf{L}_{r,i}(\beta_{r0})\}]\\ &\le \sup_{\beta_r}\log[\mathrm{E}\{\mathbf{L}_{r,i}(\beta_r)/\mathbf{L}_{r,i}(\beta_{r0})\}]=0.\end{aligned}$$

Hence $\sup_{\beta_r}\mathrm{E}[n_r^{-1}\{\sum_{i=1}^{n_r}\mathcal{L}^I_{r,i}(\beta_r)-\mathcal{L}^I_{r,i}(\beta_{r0})\}]\le 0$. By Assumption 4a and the Uniform Law of Large Numbers result, this means that almost surely as $n_r\to\infty$, $n_r^{-1}\sum_{i=1}^{n_r}\widetilde{\mathcal{L}}^I_{r,i}(\beta_r,\kappa_{r1},\kappa_{r2})<n_r^{-1}\sum_{i=1}^{n_r}\widetilde{\mathcal{L}}^I_{r,i}(\beta_{r0},\kappa_{r1},\kappa_{r2})$ when $\beta_r\neq\beta_{r0}$. Therefore for arbitrary $\epsilon>0$, and for large n_r, $n_r^{-1}\sum_{i=1}^{n_r}\widetilde{\mathcal{L}}^I_{r,i}(\beta_{r0},\kappa_{r1},\kappa_{r2})$ will exceed $n_r^{-1}\sum_{i=1}^{n_r}\widetilde{\mathcal{L}}^I_{r,i}(\beta_{r0}\pm\epsilon,\kappa_{r1},\kappa_{r2})$. By continuity of $n_r^{-1}\sum_{i=1}^{n_r}\widetilde{\mathcal{L}}^I_{r,i}(\beta_r,\kappa_{r1},\kappa_{r2})$ there must be a local maximum of this function in the interval $(\beta_{r0}-\epsilon,\beta_{r0}+\epsilon)$. Since $n_r^{-1}\sum_{i=1}^{n_r}\widetilde{\mathcal{L}}^I_{r,i}(\beta_r,\kappa_{r1},\kappa_{r2})$ is differentiable, its derivative has to be equal 0 at this point. Since ϵ is arbitrary there must be a root of PCL score equation which is consistent for β_{r0}.

We now apply the Central Limit Theorem to the functions $n_r^{-1/2}\sum_{i=1}^{n_r}\mathcal{A}^I_{r,i}(\beta_{r0})$. Therefore, in our notation, $g_i(W_i)$ is $\mathcal{A}_{r,i}(\beta_{r0})$, V_i is λ_{ri} and W_i is D_{ri}. All the conditions of the Central Limit Theorem are easily verified except the convergence of the variances. To see such convergence, note that

$$\begin{aligned}\mathrm{var}\{n_r^{-1/2}\sum_{i=1}^{n_r}\mathcal{A}^I_{r,i}(\beta_{r0})\} &= \mathrm{var}\{n_r^{-1/2}\sum_{i=1}^{n_r}v_{r,i}(\beta_{r0})\mathcal{Z}_{r,i}(\beta_{r0})\}=\Sigma^I_{nr,0}+\Sigma^I_{nr,c};\\ \Sigma^I_{nr,0} &= n_r^{-1}\sum_{i=1}^{n_r}v_{r,i}(\beta_{r0})\mathrm{var}\{\mathcal{Z}_{r,i}(\beta_{r0})\}v^{\mathrm{T}}_{r,i}(\beta_{r0})\\ \Sigma^I_{nr,c} &= n_r^{-1}\sum_{i=1}^{n_r}\sum_{j\neq i}v_{r,i}(\beta_{r0})\mathrm{cov}\{\mathcal{Z}_{r,i}(\beta_{r0}),\mathcal{Z}_{r,j}(\beta_{r0})\}v^{\mathrm{T}}_{r,j}(\beta_{r0}).\end{aligned}$$

Using Assumptions 1 and 2 it is easy to see that $\Sigma^I_{nr,0}$ converges to Σ^I_{r0} defined as

$$\Sigma^I_{r0} \;=\; \int_{\widetilde{x}} \phi\{\mu_r(\beta_{r0},\widetilde{x})\mathcal{Y}_r(\beta_{r0},\widetilde{x})\mathcal{V}_r(\beta_{r0},\widetilde{x})\mathcal{V}_r^{\mathrm{T}}(\beta_{r0},\widetilde{x})dF_{G,r}(\widetilde{x}), \tag{B.2}$$

where in (B.2) we are using the notation such as $\mu_{ri}(\beta_{r0}) = \mu_r(\beta_{r0},\widetilde{X}_{ri})$. The limit exists because the arguments in the integral are continuous and bounded in $\widetilde{x}$. By Assumption 5, Σ^I_{r0} is positive definite. Hence $\Sigma^I_{nr,0}+\Sigma^I_{nr,c} \geq c_1\mathrm{I}$, where $c_1 > 0$. Thus we can apply the CLT to conclude that $(\Sigma^I_{nr,0}+\Sigma^I_{nr,c})^{-1/2} n_r^{-1/2}\sum_i \mathcal{A}^I_{r,i}(\beta_{r0})$ converges in distribution to Normal$(0,\mathrm{I})$.

Moreover, using Assumption 5 and results of White (1984, p.47, Theorem 3.49) and Kolmogorov and Rozanov (1960, Theorems 1 and 2), it follows that for any $i \neq j$, $\mathrm{cov}\{\mathcal{Z}_{r,i}(\beta_{r0}),\mathcal{Z}_{r,j}(\beta_{r0})\} = \mathrm{O}\{|i-j|^{-M}\}$, where $M > 1$. Thus, for some $c_2 < \infty$, for any i

$$\sum_{j\neq i} v_{r,i}(\beta_{r0})\mathrm{cov}\{\mathcal{Z}_{r,i}(\beta_{r0}),\mathcal{Z}_{r,j}(\beta_{r0})\}v^{\mathrm{T}}_{r,j}(\beta_{r0}) < c_2 < \infty.$$

Hence $\mathrm{var}\{n_r^{-1/2}\sum_{i=1}^{n_r}\mathcal{A}_{r,i}(\beta_{r0})\} < \infty$ and $\Sigma^I_{nr,c}$ is defined as

$$\Sigma^I_{nr,c} \;=\; n_r^{-1}\sum_{i=1}^{n_r}\sum_{j\neq i}^{n_r}\mathcal{H}_{r,2}(\widetilde{X}_{ri},\widetilde{X}_{rj},\beta_{r0}), \tag{B.3}$$

where

$$\mathcal{H}_{r,2}(\widetilde{x}_1,\widetilde{x}_2,\beta_{r0}) = \mathcal{Y}_{ri}(\beta_{r0})\mathcal{Y}_{rj}(\beta_{r0})\{\mathcal{V}_r(\beta_{r0},\widetilde{x}_1)\}^T\mathcal{V}_r(\beta_{r0},\widetilde{x}_2)$$
$$\times[\Phi_2\{\mu_r(\beta_{r0},\widetilde{x}_1),\mu_r(\beta_{r0},\widetilde{x}_2),\sigma^2_\lambda\Omega_{r12}(\rho)/(1+\sigma^2_\lambda)\} - \Phi\{\mu_r(\beta_{r0},\widetilde{x}_1)\}\Phi\{\mu_r(\beta_{r0},\widetilde{x}_2)\}],$$

where $\Omega_{r12}(\rho)$ is the correlation for the locations $\widetilde{x}_1$ and $\widetilde{x}_2$.

Now remember that in Theorem 1, $\kappa_{rj} = c_{rj}n_r^{-\nu}$ where $\nu \geq 1/2$. By a Taylor expansion,

$$0 \;=\; n_r^{-1/2}\sum_{i=1}^{n_r}\mathcal{A}^I_{r,i}(\widehat{\beta}^I_r) - n_r^{-\nu+1/2}M(c_{r1},c_{r2})\widehat{\beta}^I_r$$

$$= \quad n_r^{-1/2}\sum_{i=1}^{n_r}\mathcal{A}^I_{r,i}(\beta_{r0}) - n_r^{-\nu+1/2}M(c_{r1},c_{r2})\beta_{r0}$$
$$+\{n_r^{-1}\sum_{i=1}^{n_r}\frac{\partial\mathcal{A}^I_{r,i}(\widetilde{\beta}_r)}{\partial\beta_r} - n_r^{-\nu}M(c_{r1},c_{r2})\}n_r^{1/2}(\widehat{\beta}^I_r-\beta_{r0}) + o_p(1), \quad \text{(B.4)}$$

where $\widetilde{\beta}_r$ lies between $\widehat{\beta}^I_r$ and β_{r0}. Note that $\widetilde{\beta}_r$ is consistent for β_{r0}. using this fact, the uniform law of large numbers and the fact that the first order composite loglikelihood for an individual observation is a legitimate marginal likelihood, it follows that

$$-\{n_r^{-1}\sum_{i=1}^{n_r}\frac{\partial\mathcal{A}^I_{r,i}(\widetilde{\beta}_r)}{\partial\beta_r} - n_r^{-\nu}M(c_{r1},c_{r2})\} - (\Sigma^I_{nr,0}) = o_p(1),$$

and hence that

$$\Sigma^I_{nr,0}n_r^{1/2}(\widehat{\beta}^I_r-\beta_{r0}) = n_r^{-1/2}\sum_{i=1}^{n_r}\mathcal{A}^I_{r,i}(\beta_{r0}) - n_r^{-\nu+1/2}M(c_{r1},c_{r2})\beta_{r0} + o_p(1),$$

from which Theorem 1 follows immediately with

$$\Delta^I_{nr} = n_r^{-\nu+1/2}(\Sigma^I_{nr,0}+\Sigma^I_{nr,c})^{-1}M(c_{r1},c_{r2}); \quad \text{(B.5)}$$
$$\Sigma^I_{nr} = (\Sigma^I_{nr,0})^{-1}(\Sigma^I_{nr,0}+\Sigma^I_{nr,c})(\Sigma^I_{nr,0})^{-1}. \quad \text{(B.6)}$$

The proof of Theorem 2 is much the same as that for Theorem 1, and we will not repeat the details. See Section 2 for the definitions of Θ and $M^{II}(\widetilde{\kappa}_1,\widetilde{\kappa}_2,\kappa_3)$.

Define $Y_{r,ij} = (Z^{(0,0)}_{r,ij}, Z^{(0,1)}_{r,ij}, Z^{(1,0)}_{r,ij}, Z^{(1,1)}_{r,ij})^{\mathrm{T}}$ and $U(\widetilde{X}_{ri},\widetilde{X}_{rj},\Theta) = \partial[\log\{\pi^{0,0}_{r,ij}(\Theta)\}, \log\{\pi^{0,1}_{r,ij}(\Theta)\}, \log\{\pi^{1,0}_{r,ij}(\Theta)\}, \log\{\pi^{1,1}_{r,ij}(\Theta)\}]/\partial\Theta$. Then we can write the derivative of the score as $\partial\mathcal{L}^{II}_{r,ij}(\Theta)/\partial\Theta = U(\widetilde{X}_{ri},\widetilde{X}_{rj},\Theta)Y_{r,ij}$. Let $N = \min_r n_r$. In this notation, the estimator $\widehat{\Theta}$ satisfies

$$0 = \mathcal{W}^{-1}\sum_{r=1}^{R}\sum_{i,j=1}^{n_r} w_{rij}U(\widetilde{X}_{ri},\widetilde{X}_{rj},\Theta)Y_{r,ij} - \mathcal{W}^{-\nu}M^{II}(\widetilde{c}_1,\widetilde{c}_2,c_3)\Theta.$$

For $(i,j)\neq(k,\ell)$, $\mathrm{cov}(Y_{r,ij},Y_{r,k\ell}) = \Sigma_{y2}(\widetilde{X}_{ri},\widetilde{X}_{rj},\widetilde{X}_{rk},\widetilde{X}_{r\ell},\Theta)$, and also write $\mathrm{cov}(Y_{r,ij}) =$

$\Sigma_{y1}(\widetilde{X}_{ri}, \widetilde{X}_{rj}, \Theta)$. The analogue to (B.2) is

$$\Sigma^{II}_{\mathcal{W},0} = \mathcal{W}^{-1} \sum_{r=1}^{R} \sum_{i,j=1}^{n_r} w^2_{rij} U(\widetilde{X}_{ri}, \widetilde{X}_{rj}, \Theta_0) \Sigma_{y1}(\widetilde{X}_{ri}, \widetilde{X}_{rj}, \Theta_0) U(\widetilde{X}_{ri}, \widetilde{X}_{rj}, \Theta_0)^{\mathrm{T}}.$$

The analogue to (B.3) is

$$\begin{aligned} \Sigma^{II}_{\mathcal{W},c} &= \mathcal{W}^{-1} \sum_{r=1}^{R} \sum_{(i,j)\neq(k,\ell)=1}^{n_r} w_{rij} w_{rk\ell} U(\widetilde{X}_{ri}, \widetilde{X}_{rj}, \Theta_0) \Sigma_{y2}(\widetilde{X}_{ri}, \widetilde{X}_{rj}, \widetilde{X}_{rk}, \widetilde{X}_{r\ell}, \Theta_0) \\ &\quad \times U(\widetilde{X}_{rk}, \widetilde{X}_{r\ell}, \Theta_0). \end{aligned}$$

Then

$$\Sigma^{II}_{\mathcal{W}} = (\Sigma^{II}_{\mathcal{W},0})^{-1} (\Sigma^{II}_{\mathcal{W},0} + \Sigma^{II}_{\mathcal{W},c}) (\Sigma^{II}_{\mathcal{W}0})^{-1}. \tag{B.7}$$

In addition, it follows that

$$\Delta^{II}_{\mathcal{W}} = \mathcal{W}^{-\nu+1/2} (\Sigma^{II}_{\mathcal{W},0} + \Sigma^{II}_{\mathcal{W},c})^{-1} M^{II}(\widetilde{c}_1, \widetilde{c}_2, c_3). \tag{B.8}$$

We showed that

$$(\Sigma^{I}_{nr,0}) \mathcal{W}^{-1/2} (\widehat{\beta}^{I}_r - \beta_{r0}) = (n_r/\mathcal{W})^{1/2} n_r^{-1/2} \sum_{i=1}^{n_r} A^I_r(\widetilde{X}_{ri}, \beta_{r0}) + o_p(1).$$

Let $u(\widetilde{X}_{ri}, \widetilde{X}_{rj}, \Theta)$ be a vector of partial derivatives of $\log\{\pi^{kl}_{r,ij}(\Theta)\}$ with respect to θ. Then the two-stage estimator of θ, $\widehat{\theta}^{II*}$, satisfies

$$0 = \mathcal{W}^{-1} \sum_{r=1}^{R} \sum_{i,j=1}^{n_r} w_{rij} u(\widetilde{X}_{ri}, \widetilde{X}_{rj}, \widehat{\mathcal{B}}, \theta) Y_{r,ij} - \mathcal{W}^{-\nu_\theta} M^{II*}(c_3)\theta$$

Let $\Sigma_{yz}(\widetilde{X}_{ri}, \widetilde{X}_{rj}, \widetilde{X}_{rk}, \Theta) = \mathrm{cov}(Y_{r,ij}, \mathcal{Z}_{r,k})$. Matrix $\Sigma^{II}_{\mathcal{W},0}$ can be partitioned as follows

$$\begin{pmatrix} \Sigma^{II,\mathcal{B}}_{\mathcal{W},0} & (\Sigma^{II,\theta\mathcal{B}}_{\mathcal{W},0})^T \\ \Sigma^{II,\theta\mathcal{B}}_{\mathcal{W},0} & \Sigma^{II,\theta}_{\mathcal{W},0} \end{pmatrix}$$

according to parameters $(\mathcal{B}, \theta)$. Matrix $\Sigma^{II}_{\mathcal{W},c}$ can be partitioned similarly on $\Sigma^{II,\mathcal{B}}_{\mathcal{W},c}$,

$\Sigma^{II,\theta\mathcal{B}}_{\mathcal{W},c}$ and $\Sigma^{II,\theta}_{\mathcal{W},c}$. The analogue to (B.2) is the part of $\Sigma^{II}_{\mathcal{W},0}$ corresponding to covariance parameters θ, $\Sigma^{II,\theta}_{\mathcal{W},0}$. The analogue to (B.3) is the sum of $\Sigma^{II,\theta}_{\mathcal{W},c}$ and $\Sigma^{II*}_{\mathcal{W},c}$, where $\Sigma^{II*}_{\mathcal{W},c}$ equals to

$$\Sigma^{II*}_{\mathcal{W},c} = \sum_{r=1}^{R} \{\Sigma^{II,\theta\beta_r}_{\mathcal{W},0} \Sigma^{I}_{n,r} - \Sigma^{II,I}_{\mathcal{W},r} (\Sigma^{I}_{nr,0})^{-1}\} (\Sigma^{II,\theta\beta_r}_{\mathcal{W},0})^{T},$$

where $\Sigma^{II,\theta\beta_r}_{\mathcal{W},0}$ is part of the further partitioned matrix $\Sigma^{II,\theta\mathcal{B}}_{\mathcal{W},0} = [\Sigma^{II,\theta\beta_1}_{\mathcal{W},0}, ..., \Sigma^{II,\theta\beta_r}_{\mathcal{W},0}]$ and

$$\Sigma^{II,I}_{\mathcal{W},r} = \mathcal{W}^{-1} \sum_{i,j=1}^{n_r} \sum_{k=1}^{n_r} w_{rij} u(\widetilde{X}_{ri}, \widetilde{X}_{rj}, \Theta_0) \Sigma_{yz}(\widetilde{X}_{ri}, \widetilde{X}_{rj}, \widetilde{X}_{rk}, \Theta_0) \{v_{r,k}(\beta_{r0})\}^{T}.$$

Then

$$\Sigma^{II*}_{\mathcal{W}} = (\Sigma^{II,\theta}_{\mathcal{W},0})^{-1} (\Sigma^{II,\theta}_{\mathcal{W},0} + \Sigma^{II,\theta}_{\mathcal{W},c} + \Sigma^{II*}_{\mathcal{W},c}) (\Sigma^{II,\theta}_{\mathcal{W},0})^{-1}. \tag{B.9}$$

In addition if follows that

$$\Delta^{II*}_{\mathcal{W}} = (\Sigma^{II,\theta}_{\mathcal{W},0})^{-1} [- \sum_{r=1}^{R} \Sigma^{II,\theta\beta_r\Delta_r}_{\mathcal{W},0}, \mathcal{W}^{-\nu_\theta+1/2} M^{II*}(c_3)]. \tag{B.10}$$

The main purpose of this section is to show that smoothing parameters should be of order $O(n_r^{-1})$, and to sketch the basic algebra to obtain these smoothing parameters. We describe the results only for the composite likelihood of the first order and for the two-stage method: the composite likelihood of the second order is similar but more algebraically intense. It is worth remembering our notation. Define $\mathcal{Z}_{r,i}(\beta_r) = D_{ri} - \Phi\{\mu_{ri}(\beta_r)\}$. Let $\mathcal{Y}_{ri}(\beta_r) = \phi\{\mu_{ri}(\beta_r)\}(\Phi\{\mu_{ri}(\beta_r)\}[1 - \Phi\{\mu_{ri}(\beta_r)\}])^{-1}$. Write $\partial\mu_{ri}(\beta_r)/\partial\beta_r = \mathcal{V}_{ri}(\beta_r)$, and write $v_{r,i}(\beta_r) = \mathcal{V}_{ri}(\beta_r)\mathcal{Y}_{ri}(\beta_r)$. The first derivative of $\mathcal{L}^{I}_{r,i}(\beta_r)$ is $\mathcal{A}^{I}_{r,i}(\beta_r) = v_{r,i}(\beta_r)\mathcal{Z}_{r,i}(\beta_r)$. The derivative of $\mathcal{A}^{I}_{r,i}(\beta_r)$ with respect to β_r and evaluated at β_{r0} is easily seen to have expectation $\mathcal{N}_{ri} = -\phi\{\mu_r(\beta_{r0}, \widetilde{X}_{r,i})\mathcal{Y}_r(\beta_{r0}, \widetilde{X}_{ri})\mathcal{V}_r(\beta_{r0}, \widetilde{X}_{ri})\mathcal{V}_r^{\mathrm{T}}(\beta_{r0}, \widetilde{X}_{ri})$, which is minus the covariance of $\mathcal{A}^{I}_{r,i}(\beta_{r0})$. Also, let $\varpi_{2,ri}(\beta) = -\ln[1 - \Phi\{\mu_{ri}(\beta)\}]$ and $\varpi_{1,ri}(\beta) = \varpi_{2,ri}(\beta) +$

$\ln[\Phi\{\mu_{ri}(\beta)\}]$.

Define $\mathcal{T}(\kappa_1, \kappa_2, \beta) = M(\kappa_1, \kappa_2, \beta)\beta$, where as before $M(\kappa_{r1}, \kappa_{r2}) = \text{diag}(\kappa_{r1}\mathcal{G}_{r1}$, $\kappa_{r2}\mathcal{G}_{r2}$, $0I_{q_\zeta - 1})$.

Recall that the smoothing parameters in Theorem 1 defined as $\kappa_{rj} = n_r^{-\nu}c_{rj}$, $j = 1, 2$. Assume that $\nu \geq 1/2$, then by Theorem 1 $(\widehat{\beta}_r - \beta_r) = O_p(n_r^{-1/2})$. Thus by the Taylor expansion,

$$\begin{aligned}
\Phi\{\mu_{ri}(\widehat{\beta}_r)\} &= \Phi\{\mu_{ri}(\beta_{r0})\} + \phi\{\mu_{ri}(\beta_{r0})\}\mathcal{V}_{ri}^{\mathrm{T}}(\beta_{r0})(\widehat{\beta}_r - \beta_{r0}) + O_p(n_r^{-1}); \\
\varpi_{1,ri}(\widehat{\beta}_r) &= \varpi_{1,rj}(\beta_r) + \mathcal{Y}_{ri}(\beta_{r0})\mathcal{V}_{ri}^{\mathrm{T}}(\beta_{r0})(\widehat{\beta}_r - \beta_{r0}) + O_p(n_r^{-1}).
\end{aligned}$$

Define $\mathcal{Q}_{ri} = [\phi\{\mu_{ri}(\beta_{r0})\}\mathcal{Y}_{ri}(\beta_{r0})]^{1/2}\mathcal{V}_{ri}(\beta_{r0})$. Thus $\text{SKL} = n_r^{-1}\sum_{i=1}^{n_r}\mathcal{Q}_{ri}^{\mathrm{T}}(\widehat{\beta}_r - \beta_{r0})(\widehat{\beta}_r - \beta_{r0})^{\mathrm{T}}\mathcal{Q}_{ri} + O_p(n_r^{-3/2})$, where It is easy to see that $\text{SKL} = n_r^{-1}\sum_j \mathcal{Q}_{rj}^{\mathrm{T}}(\widehat{\beta}_r^* - \beta_{r0})(\widehat{\beta}_r^* - \beta_{r0})^{\mathrm{T}}\mathcal{Q}_{rj} + O_p(n_r^{-3/2})$, where $\widehat{\beta}_r^* = \beta_{r0} + \{\Sigma_{r0}^I + M(\kappa_{r1}, \kappa_{r2})\}^{-1}\{n_r^{-1}\sum_j \mathcal{A}_{r,j}(\beta_r) - \mathcal{T}(\kappa_{r1}, \kappa_{r2}, \beta_r)\}$. The associated with SKL is its expected value in this asymptotically equivalent version of $\widehat{\beta}_r$, namely

$$\text{MASKL} = n_r^{-1}\sum_j \mathcal{Q}_{rj}^{\mathrm{T}}\mathrm{E}\{(\widehat{\beta}_r^* - \beta_r)(\widehat{\beta}_r^* - \beta_r)^{\mathrm{T}}\}\mathcal{Q}_{rj} + \mathrm{E}\{O_p(n_r^{-3/2})\}.$$

Note that $n_r^{-1}\sum_{i=1}^{n_r}\mathcal{Q}_{ri}\mathcal{Q}_{ri}^{\mathrm{T}} = \Sigma_{nr,0}^I + O_p(n_r^{-1/2})$. Define $R(\kappa_{r1}, \kappa_{r2}) = \{\Sigma_{nr,0}^I + M(\kappa_{r1}, \kappa_{r2})\}^{-1}$, $S(\kappa_{r1}, \kappa_{r2}) = n_r^{-1}V_r + \mathcal{T}(\kappa_{r1}, \kappa_{r2}, \beta_{r0})\mathcal{T}^{\mathrm{T}}(\kappa_{r1}, \kappa_{r2}, \beta_{r0})$ and $V_r = \text{cov}\{n_r^{-1/2}\sum_i \mathcal{A}_{r,i}(\beta_{r0})\}$. Then we have that the asymptotically equivalent version is

$$\widetilde{\text{MASKL}}(\kappa_{r1}, \kappa_{r2}) = \text{trace}\{R(\kappa_{r1}, \kappa_{r2})S(\kappa_{r1}, \kappa_{r2})R(\kappa_{r1}, \kappa_{r2})\Sigma_{nr,0}^I\}. \quad \text{(B.11)}$$

It is evident by inspection that in order to minimize (B.11), we must have $(\kappa_{r1}, \kappa_{r2}) = O(n_r^{-1})$ as claimed. The minimizer of (B.11) solves $\partial\widetilde{\text{MASKL}}(\kappa_{r1}, \kappa_{r2})/\partial(\kappa_{r1}, \kappa_{r2})$. Then there is a 2×2 matrix $\mathcal{H}_r(\widetilde{\kappa}^{(r)}, \beta_{r0}, \Sigma_{n,r})$ and a 2×1 vector $\mathcal{J}_r(\widetilde{\kappa}^{(r)}, \beta_{r0}, \Sigma_{n,r})$, both of order $O(1)$, such that to terms of order $o(n_r^{-1})$, the nontrivial equation for

$\kappa^{(r)}$ is $\widetilde{\kappa}^{(r)} = n_r^{-1}\mathcal{H}_r^{-1}(\widetilde{\kappa}^{(r)}, \beta_{r0}, \Sigma_{n,r})\mathcal{J}_r(\widetilde{\kappa}^{(r)}, \beta_{r0}, \Sigma_{n,r})$=0. The later equation can be solved iteratively. However, to terms of first order, the minimizer of (B.11) solves

$$0 = \left\{\frac{\partial \widetilde{\mathrm{MASKL}}(0,0)}{\partial(\kappa_{r1}, \kappa_{r2})^{\mathrm{T}}}\right\} + \left\{\frac{\partial^2 \widetilde{\mathrm{MASKL}}(0,0)}{\partial(\kappa_{r1}, \kappa_{r2})^{\mathrm{T}}\partial(\kappa_{r1}, \kappa_{r2})}\right\}(\kappa_{r1}, \kappa_{r2})^{\mathrm{T}},$$

which leads to a linear question $\widetilde{\kappa}^{(r)} = n_r^{-1}\widetilde{\mathcal{H}}_r^{-1}(\beta_{r0}, \Sigma_{n,r})\widetilde{\mathcal{J}}_r(\beta_{r0}, \Sigma_{n,r})$. All terms mentioned above are straightforward to compute and estimate. In the case that one sets all the smoothing parameters equal for $r = 1, .., R$, one simply replace (B.11) by its sum over R.

Similar to the previous results, the expected value of SKL using asymptotically equivalent version of $\widehat{\theta}$, equals to

$$\mathrm{MASKL}^{II*} = \mathcal{W}^{-1}\sum_r\sum_{i,j}(\mathcal{Q}_{rij}^{II*})^{\mathrm{T}}\mathrm{E}\{(\widehat{\theta}^* - \theta)(\widehat{\theta}^* - \theta)^{\mathrm{T}}\}\mathcal{Q}_{rij}^{II*} + \mathrm{E}\{O_p(\mathcal{W}^{-3/2})\},$$

where $\mathcal{W}^{-1}\sum_r\sum_{i,j}\mathcal{Q}_{rij}^{II*}(\mathcal{Q}_{rij}^{II*})^{\mathrm{T}} = \Sigma_{\mathcal{W},0}^{II,\theta} + O_p(W^{-1/2})$. Define $R^{II*}(\kappa_3) = \{\Sigma_{\mathcal{W},0}^{II,\theta} + M^{II*}(\kappa_3)\}^{-1}$, $S^{II*}(\kappa_3) = \mathcal{W}^{-1}V^{II*} + \mathcal{T}^{II*}(\kappa_3, \theta_0)\{\mathcal{T}^{II*}(\kappa_3, \theta_0)\}^{\mathrm{T}}$, where $\mathcal{T}^{II*}(\kappa_3, \theta_0) = M^{II*}(\kappa_3)\theta_0$ and $V^{II*} = \mathrm{cov}\{\mathcal{W}^{-1/2}\sum_r\sum_{ij}\partial\mathcal{L}_{r,ij}^{II}(\widehat{\beta}_r, \theta_0)/\partial\theta\}$. Then the asymptotically equivalent version of MASKL^{II*} is

$$\widetilde{\mathrm{MASKL}}^{II*}(\kappa_3) = \mathrm{trace}\{R^{II*}(\kappa_3)S^{II*}(\kappa_3)R^{II*}(\kappa_3)\Sigma_{\mathcal{W},0}^{II,\theta}\}. \tag{B.12}$$

It is clear that $\kappa_3 = O(\mathcal{W}^{-1})$ as claimed. To terms of first order, the minimizer of (B.12) solves

$$0 = \left\{\frac{\partial \widetilde{\mathrm{MASKL}}^{II*}(0)}{\partial\kappa_3^{\mathrm{T}}}\right\} + \left\{\frac{\partial^2 \widetilde{\mathrm{MASKL}}^{II*}(0)}{\partial\kappa_3^{\mathrm{T}}\partial\kappa_3}\right\}\kappa_3^{\mathrm{T}}.$$

These terms are straightforward to compute and estimate.

VITA

Tatiyana(Tanya) V. Apanasovich was born in Minsk, Belarus. She is the first daughter of Vladimir Vladimirovich Apanasovich and Tamara Vladimirovna Apanasovich. Tatiyana graduated from Belarusian State University in Minsk, Belarus in May 1999 with a Bachelor of Science degree in applied mathematics. That same year, she was admitted to the Ph.D program in the Department of Statistics at Texas A&M University. Tatiyana received her Ph.D degree in August 2004.

Her permanent address is

59/2-144 Gazeta Izvestia Ave.

Minsk, Belarus