Biometrika (2016), *pp.* 1–15
Printed in Great Britain
doi: 10.1093/biomet/asw003

Regression analysis of networked data

BY YAN ZHOU AND PETER X.-K. SONG

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
zhouyan@umich.edu pxsong@umich.edu

SUMMARY

This paper concerns regression methodology for assessing relationships between multi-dimensional response variables and covariates that are correlated within a network. To address analytical challenges associated with the integration of network topology into the regression analysis, we propose a hybrid quadratic inference method that uses both prior and data-driven correlations among network nodes. A Godambe information-based tuning strategy is developed to allocate weights between the prior and data-driven network structures, so the estimator is efficient. The proposed method is conceptually simple and computationally fast, and has appealing large-sample properties. It is evaluated by simulation, and its application is illustrated using neuroimaging data from an association study of the effects of iron deficiency on auditory recognition memory in infants.

Some key words: Estimating function; Event-related potential; Generalized method of moments; Hybrid quadratic inference function; Shrinkage.

1. INTRODUCTION

Data collected from networks are common in practice. A network refers to a set of nodes or vertices which are joined in pairs by edges (Newman, 2010). An important feature of a network is that, unlike in a space-time system, between-node distance may not be defined precisely by a numerical metric. In this paper we discuss regression analysis of multi-dimensional response variables on covariates that are collected from networks. Although considerable attention has been paid to methods of learning network topology, little work has been done on regression, which plays a central role in the study of response-covariate relationships. Because data from a network are correlated across nodes, to achieve high statistical efficiency one needs to incorporate appropriate dependence structures into inference, an issue that we address here.

Networked data have more complex dependence mechanisms than can be described by conventional covariance or correlation matrices. For example, dependence symmetry among nodes may not hold, and it may not be possible to model strength of dependence explicitly due to the lack of a legitimate distance function. Our motivating example comes from a project, in collaboration with scientists at the Center for Human Growth and Development of the University of Michigan, whose scientific objective is to evaluate whether iron deficiency affects auditory recognition memory in infants and, if so, how. An infant's memory capability is measured by electrical activity in the brain during a period of 2000 milliseconds using an electroencephalography, EEG, net consisting of 64-channel sensors on the scalp; see Fig. 1(a).

The data are collected at two times: when an infant hears his or her mother's voice and when he or she hears a stranger's voice. At each time, three event-related potentials, P2, P750 and late slow wave, are recorded after standard data processing. These three event-related potentials are

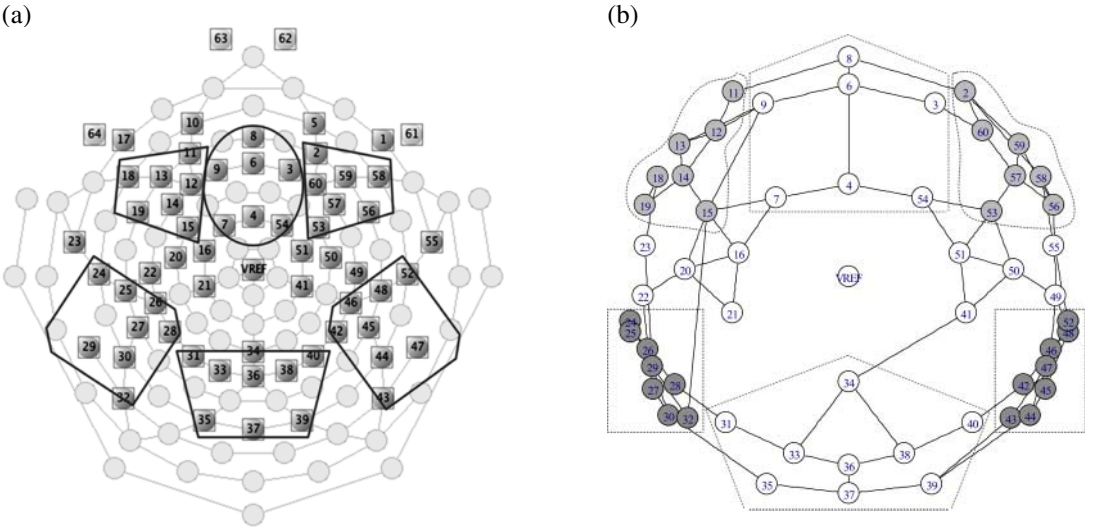

Fig. 1. (a) Layout of the 64-channel sensor net, where the six outlined clusters of nodes relate to auditory recognition memory and the remaining nodes belong to an additional cluster. (b) Sparse graphical representation of the learned network among electrodes based on the late slow wave data under voice stimulus from a stranger.

widely used as primary outcomes of auditory recognition memory (Siddappa et al., 2004; Mai et al., 2012). In this paper we consider only the late slow wave outcome. Such measurements from the 64 electrodes are correlated in the EEG net, and the correlation is highly clustered according to subregions of memory functionality. Correlations of late slow wave measurements are not necessarily symmetric over the 64 nodes. Standard analysis of the event-related potential data using spatial analysis-of-variance mixed-effects models (Gevins & Smith, 2000; Fields & Kuperberg, 2012) assumes implicitly symmetric exchangeable correlations among the 64 nodes for late slow wave data, and fails to detect significant association between iron deficiency and late slow wave activity.

To improve upon the standard analysis, we treat the EEG net as a network and develop a flexible dependence model that can better reflect the underlying relationships among the electrodes, for instance allowing for clustered and asymmetric dependence relationships. In particular, we develop a strategy to combine two sources of knowledge concerning the network topology: our collaborators' expertise regarding established or prior knowledge about subregions of memory functionality, and dependencies learned from data. Some popular statistical methods that have been used to learn sparse conditional dependence structures of networks include: sparse partial correlation (Peng et al., 2009), implemented in the R (R Development Core Team, 2016) package space; the graphical lasso (Yuan & Lin, 2007), implemented in the R package glasso; neighbourhood selection (Meinshausen & Buehlmann, 2006), also implemented in the R package glasso; and the sparse joint additive model (Voorman et al., 2014), implemented in the R package spacejam.

We consider marginal regression for networked data, which allows various forms of dependence among nodes and can easily handle categorical outcomes. For estimation of regression coefficients in the marginal model, both generalized estimating equations (Liang & Zeger, 1986) and quadratic inference functions (Qu et al., 2000) have been extensively studied. However, these methods cannot be applied directly to networked data because of challenges in incorporating network dependence structures. One desirable method for fitting the marginal model under unstructured correlation is the adaptive estimating equation method of Qu & Lindsay (2003), which does

not require the inverse of a correlation matrix. A disadvantage of using unstructured correlation in generalized estimating equations or Qu & Lindsay's adaptive quadratic inference function is the involvement of a large number of nuisance parameters in the estimation, leading to potential loss of estimation efficiency and numerical instability. Many authors have advocated incorporating correlation structures to achieve good estimation efficiency; see, for example, Pan (2001), Qu et al. (2008) and Zhou & Qu (2012).

Our strategy of combining two sources of network topology follows the linear shrinkage estimation approach of Stein (1956), which is discussed by Ledoit & Wolf (2004) in the context of covariance matrix estimation. We propose to shrink an unstructured covariance matrix towards a prior or target network structure, represented by an adjacency matrix with elements 0 representing no connection and elements 1 representing the existence of a connection between nodes. Following Hansen (1982), we construct an over-identified estimating function with a shrinkage tuning parameter determined by minimizing the inverse of the Godambe information. Our estimation method allocates higher weights to more relevant correlation structures while down-weighting others. The process of tuning does not affect estimation consistency or asymptotic normality but gains efficiency when done properly.

2. Framework

2·1. *Estimating functions*

Suppose that the response variable y_{ij} and the associated p-dimensional covariate x_{ij} are measured at node, or vertex, j for subject i ($j = 1, \ldots, m; i = 1, \ldots, n$). Let $y_i = (y_{i1}, \ldots, y_{im})^{\mathrm{T}}$ and $x_i = (x_{i1}, \ldots, x_{im})^{\mathrm{T}}$, which is an $m \times p$ matrix, and let (y_i, x_i) ($i = 1, \ldots, n$) be independent and identically distributed data from n subjects. To perform a regression analysis of the networked data, we adopt a population-average model framework with mean model $\mu_{ij} = E(y_{ij} \mid x_{ij}) = \mu(x_{ij}^{\mathrm{T}}\beta)$, where $\mu(\cdot)$ is a known link function, β is a p-dimensional parameter vector of interest, and $\mu_i = (\mu_{i1}, \ldots, \mu_{im})^{\mathrm{T}}$.

To proceed with the quasilikelihood approach to inference on β, according to Liang & Zeger (1986), the second moment of y_i is specified by $V_i = A_i^{1/2} R(\alpha) A_i^{1/2}$, with $R(\alpha)$ a working correlation matrix and A_i the diagonal matrix of marginal variances $\mathrm{var}(y_{ij} \mid x_{ij}) = \phi v(\mu_{ij})$, where $v(\cdot)$ is the variance function and ϕ the dispersion parameter. Generalized estimating equations (Liang & Zeger, 1986) provide an estimate of β by solving the equation $\sum_{i=1}^{n} \dot{\mu}_i^{\mathrm{T}} V_i^{-1} (y_i - \mu_i) = 0$, where $\dot{\mu}_i(\cdot)$ is the gradient vector of $\mu_i(\cdot)$ with respect to β; see Song (2007, Ch. 2 and 5). Because the number of nodes in a network is fixed, we write the variance V_i as simply V. Under regularity conditions, the resulting generalized estimating equations estimator is consistent and asymptotically normal, but may have low efficiency if the working correlation $R(\alpha)$ does not represent the true correlation structure adequately enough.

Many strategies have been proposed to improve the efficiency of generalized estimating equations estimators. A popular approach is the quadratic inference function procedure of Qu et al. (2000), which assumes that the inverse of the working correlation matrix, R^{-1}, may be expanded approximately as a linear combination of basis matrices,

$$R^{-1}(\alpha) = \sum_{k=0}^{K} a_k M_k, \tag{1}$$

where M_0 is the identity matrix, M_k ($k = 1, \ldots, K$) are known symmetric basis matrices with elements equal to either 0 or 1, and the a_k are unknown coefficients that may depend on the

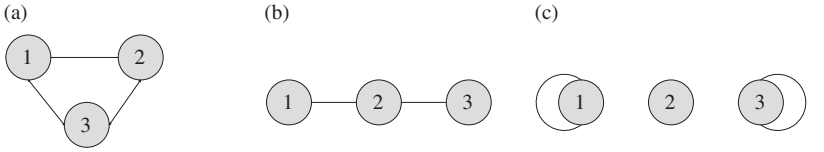

Fig. 2. Graphical display of basis matrices (a) M_{comp} (M_1), (b) M_{chain} (M_1^*), and (c) M_2^* for a three-node network.

parameter α. Then, the generalized estimating equations may be written as a linear combination of estimating functions given by the extended score vector

$$\bar{q}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} q_i(\beta) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \dot{\mu}_i^{\mathrm{T}} A_i^{-1}(y_i - \mu_i) \\ \dot{\mu}_i^{\mathrm{T}} A_i^{-1/2} M_1 A_i^{-1/2}(y_i - \mu_i) \\ \vdots \\ \dot{\mu}_i^{\mathrm{T}} A_i^{-1/2} M_K A_i^{-1/2}(y_i - \mu_i) \end{pmatrix}, \tag{2}$$

where the dimension of $\bar{q}_n(\beta)$ is $p(K+1)$. Unlike generalized estimating equations, the quadratic inference function does not require estimation of the nuisance parameter α. Because $\bar{q}_n(\beta)$ is an over-identified score vector, the equation $\bar{q}_n(\beta) = 0$ has no solution. Instead, similar to generalized method of moments (Hansen, 1982), the quadratic inference function method minimizes a quadratic objective function of the form

$$n\bar{q}_n^{\mathrm{T}}(\beta)\Gamma^{-1}(\beta)\bar{q}_n(\beta), \tag{3}$$

where the optimal weighting matrix is $\Gamma(\beta) = \mathrm{var}\{q_i(\beta)\}$, which may be consistently estimated by the sample covariance matrix $\bar{\Gamma}_n = n^{-1}\sum_{i=1}^{n} q_i(\beta)q_i^{\mathrm{T}}(\beta)$. In implementation, we adopt the unique Moore–Penrose generalized inverse in (3) to ensure numerical stability, as the matrix $\bar{\Gamma}_n$ may be singular (Hu & Song, 2012).

2·2. *Graphical interpretation of basis matrices*

We now present some geometric insights into the connection between basis matrices and network topology, using two popular correlation structures to illustrate how knowledge of the network topology may aid estimation. For ease of discussion, consider a three-dimensional network. The first example is the exchangeable correlation matrix, which according to Qu et al. (2000) has two basis matrices: $M_0 = I$, and M_1 which has 0 on the diagonal and 1 elsewhere. The other example is the first-order autoregressive, or AR(1), correlation, which has three basis matrices: $M_0 = I$, M_1^* which has 1 on the subdiagonals and 0 elsewhere, and M_2^* which has 1 in the two corner entries of the diagonal and 0 elsewhere.

These basis matrices may be viewed as adjacency matrices with the corresponding graphical representations displayed in Fig. 2. Matrix $M_0 = I$ corresponds to the adjacency matrix of an independence graph in which all nodes are disconnected. The basis matrix M_1 in Fig. 2(a) from the exchangeable correlation gives the adjacency matrix of a complete graph, denoted by M_{comp}. For the two basis matrices of the AR(1) correlation, M_1^* in Fig. 2(b) represents the adjacency matrix of a chain graph, denoted by M_{chain}, and the other matrix M_2^* in Fig. 2(c) indicates that both the beginning and end nodes are absorbing in a chain graph. Such graphical representation of between-node connectivity is a typical form of network topology knowledge available

from scientists or from a network learned by inverting the correlation matrix obtained from training or pilot study data. In the framework of quadratic inference function theory, it is feasible to incorporate adjacency matrices in inference via equation (2) for the parameters in regression models. The key insight is that each nonzero off-diagonal element in the adjacency, or basis, matrix corresponds to an edge in a graphical model that describes the existence of conditional dependence between two nodes given the other nodes. Since no numerical value for connection strength is available in an adjacency matrix, such a matrix is particularly suitable for representing prior knowledge about a network topology. In the case of exchangeable correlation, the complete network adjacency matrix M_{comp} is regarded as being sufficient, since the inverse of the correlation matrix, $R^{-1}(\alpha)$, can be fully represented by basis matrices I and M_{comp}. In the case of AR(1), the chain network adjacency matrix M_{chain} is partially sufficient, since it captures only the conditional dependence between nodes without self-connectivity of the beginning and end nodes.

2·3. *Data-driven network topology*

The quadratic inference function method may be generalized to networked data analysis if the adjacency matrices are constructed in a reasonable manner. In practice, however, the underlying graphical structures from the networked data are so complex that simple structures, such as the complete graph in Fig. 2(a) and the chain graph in Fig. 2(b), are insufficient. Using the available data, we can establish some data-driven knowledge via, for example, an unstructured dependency in which all variances and covariances are estimated. A drawback of this approach is that in a high-dimensional network, the inverse of the estimated covariance matrix could be computationally unstable or prohibitively expensive to compute by standard software. One solution given by Qu & Lindsay (2003) is the so-called adaptive procedure, which requires only estimation of the covariance matrix. It follows from the Cayley–Hamilton theorem (Bhatia, 1997) that the inverse of an $m \times m$ positive-definite matrix may be written as

$$V^{-1} = \frac{(-1)^{m-1}}{|V|}\left(c_1 I + c_2 V + \cdots + c_{m-1} V^{m-2} + V^{m-1}\right), \tag{4}$$

where c_j $(j = 1, \ldots, m-1)$ are certain suitable coefficients. Consequently, the optimal weight matrix $V^{-1}\dot{\mu}$ for a basic estimating function $s = y - \mu(\beta)$ lies in the space spanned by the columns of $\dot{\mu}, V\dot{\mu}, \ldots, V^{m-1}\dot{\mu}$. For the sake of parsimony, Qu & Lindsay (2003) suggested including in (4) only the gradient direction generated by the first two columns, $\dot{\mu}$ and $V\dot{\mu}$. This gives the extended score vector

$$\bar{h}_n(\beta) = \begin{pmatrix} \bar{h}_n^{(1)} \\ \bar{h}_n^{(2)} \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \dot{\mu}_i^{\mathrm{T}}(y_i - \mu_i) \\ \dot{\mu}_i^{\mathrm{T}} V (y_i - \mu_i) \end{pmatrix}, \tag{5}$$

where V is consistently estimated by $\hat{V} = n^{-1}\sum_{i=1}^{n} s_i s_i^{\mathrm{T}}$ with $s_i = y_i - \mu_i(\beta)$. Clearly, (5) does not require the availability of basis matrices as given in (1). However, the number of parameters to be estimated in V is large, especially in the case of complex networks, and thus overfitting may occur in determining the network dependence structure. It is therefore critical to regularize the covariance matrix estimation, so that the resulting estimated dependencies could strike a balance between parsimony and quality of fit to improve statistical power.

3. PROPOSED METHOD

3·1. *Hybrid quadratic inference function*

Inspired by the idea of shrinkage estimation (Stein, 1956), our regularization procedure involves shrinking the estimation of the covariance V towards a known prior structure Π, a given adjacency matrix, e.g., provided by an expert. We propose to construct the extended score

$$\bar{g}_n(\beta \mid \gamma) = \frac{1}{n}\sum_{i=1}^{n} g_i(\beta \mid \gamma) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \dot{\mu}_i^{\mathrm{T}} A_i^{-1}(y_i - \mu_i) \\ \dot{\mu}_i^{\mathrm{T}} \left\{ \gamma A_i^{-1/2} \Pi A_i^{-1/2} + (1-\gamma) V \right\} (y_i - \mu_i) \end{pmatrix}, \quad (6)$$

where $\gamma \in [0, 1]$ denotes the shrinkage intensity coefficient. The right-most expression in (6) is intended to provide an improvement in estimation efficiency. Let $U_i(\gamma) = \gamma A_i^{-1/2} \Pi A_i^{-1/2} + (1-\gamma)V$, a linear shrinkage estimator of V (Ledoit & Wolf, 2004). For $\gamma = 1$ the shrinkage estimator fully favours the prior target Π, whereas for $\gamma = 0$ it reduces to the unrestricted covariance V. The key feature of this approach is that it provides a systematic way to obtain a regularized dependence structure, which outperforms both $A_i^{-1/2} \Pi A_i^{-1/2}$ and V in terms of numerical stability and statistical efficiency in the estimation of β.

The extended score $\bar{g}_n$ in (6) may be rewritten as

$$\bar{g}_n(\beta \mid \gamma) = \frac{\gamma}{n}\sum_{i=1}^{n} \begin{pmatrix} \dot{\mu}_i^{\mathrm{T}} A_i^{-1}(y_i - \mu_i) \\ \dot{\mu}_i^{\mathrm{T}} A_i^{-1/2} \Pi A_i^{-1/2}(y_i - \mu_i) \end{pmatrix} + \frac{(1-\gamma)}{n}\sum_{i=1}^{n} \begin{pmatrix} \dot{\mu}_i^{\mathrm{T}} A_i^{-1}(y_i - \mu_i) \\ \dot{\mu}_i^{\mathrm{T}} V (y_i - \mu_i) \end{pmatrix}, \quad (7)$$

and so can be expressed as $\gamma \bar{f}_n(\beta \mid \Pi) + (1-\gamma)\bar{h}_n(\beta \mid V)$, where γ describes the relative weighting of importance given to $\bar{f}_n$ versus $\bar{h}_n$. We call (7) the hybrid extended score vector; it is based on unbiased estimating functions. Note that $\bar{f}_n(\beta \mid \Pi)$ can produce poor results if the target network structure Π is noninformative and far from the truth; similarly, $\bar{h}_n(\beta \mid V)$ may lose efficiency if a prior dependence structure is known but not utilized. Therefore, by allocating higher weights to more relevant extended score vectors, $\bar{g}_n$ can improve both computational performance and statistical inference for β.

Consequently, given a shrinkage coefficient γ, we can estimate β by minimizing

$$Q_n(\beta \mid \gamma) = n \bar{g}_n^{\mathrm{T}}(\beta \mid \gamma) \Gamma^{-1}(\beta \mid \gamma) \bar{g}_n(\beta \mid \gamma), \quad (8)$$

where Γ is consistently estimated by $\bar{\Gamma}_n = n^{-1}\sum_{i=1}^{n} g_i(\beta \mid \gamma) g_i^{\mathrm{T}}(\beta \mid \gamma)$. Since the estimator of β depends on the choice of shrinkage coefficient γ, it is denoted by $\hat{\beta}(\gamma)$ below.

3·2. *Asymptotic properties*

According to Hansen's theory of generalized method of moments, under certain regularity conditions (Hansen, 1982; Harris & Mátyás, 1999), the estimator of β is not only consistent but also asymptotically normally distributed. With a known target structure Π and a fixed shrinkage coefficient γ, these large-sample properties remain valid for the proposed estimator in (8). In other words, $\hat{\beta}(\gamma) \to \beta_0$ in probability as $n \to \infty$, and

$$\sqrt{n}\{\hat{\beta}(\gamma) - \beta_0\} \to N\{0, J^{-1}(\beta_0 \mid \gamma)\}$$

in distribution as $n \to \infty$, where $J(\beta_0 \mid \gamma) = G^{\mathrm{T}}(\beta_0 \mid \gamma) \Gamma^{-1}(\beta_0 \mid \gamma) G(\beta_0 \mid \gamma)$ is the Godambe information of $g_i(\beta_0 \mid \gamma)$, provided that $\bar{\Gamma}_n(\hat{\beta} \mid \gamma) \to \Gamma(\beta_0 \mid \gamma)$ in probability and $\dot{\bar{g}}_n(\hat{\beta} \mid \gamma) \to G(\beta_0 \mid \gamma)$ in probability, both of which can be routinely verified under Conditions A1–A6 in the

Appendix. The hybrid extended score vector $g_i(\beta_0 \mid \gamma)$ is constructed on the basis of a known target structure Π, so $\hat{\beta}(\gamma)$ and $J(\beta_0 \mid \gamma)$ depend not only on γ but also on Π. For notational convenience, the dependence on Π is not shown explicitly except where necessary.

In addition to the above large-sample properties, the asymptotic χ^2 distribution of the quadratic inference function (Qu et al., 2000; Qu & Lindsay, 2003) can easily be extended to (8); that is, the statistic $\hat{Q}_n\{\hat{\beta}(\gamma) \mid \gamma\}$ tends to $\chi^2_{\mathrm{rank}\{\Gamma(\beta_0|\gamma)\}-p}$ in distribution as $n \to \infty$, which is useful in testing for goodness of fit under the null hypothesis $H_0 : E(\bar{g}_n) = 0$ (Hansen, 1982). Furthermore, a generalized-method-of-moments-type test for a nested model can be derived. Consider a partition, say $\beta = \{\beta_A, \beta_B\}$, with parameter of interest β_A and nuisance parameter β_B. To test the null hypothesis $H_0 : \beta_A = a_0$, a test statistic $Q_n\{a_0, \tilde{\beta}_B(\gamma) \mid \gamma\} - Q_n\{\hat{\beta}_A(\gamma), \hat{\beta}_B(\gamma) \mid \gamma\}$, where $\tilde{\beta}_B = \arg\min_{\beta_B} Q_n(a_0, \beta_B \mid \gamma)$ and $\{\hat{\beta}_A(\gamma), \hat{\beta}_B(\gamma)\} = \arg\min_{(\beta_A,\beta_B)} Q_n(\beta_A, \beta_B \mid \gamma)$, tends to $\chi^2_{\dim(a_0)}$ in distribution as $n \to \infty$. The degrees of freedom of this asymptotic χ^2 distribution under $H_0 : \beta_A = a_0$ does not depend on γ.

3·3. *Choice of the shrinkage coefficient*

We wish to determine a shrinkage coefficient γ to find a balance between two types of network dependence structure under a certain optimality criterion. We propose to select γ by minimizing the trace of the inverse of the Godambe information matrix $J(\beta_0 \mid \gamma)$, in order to maximize estimation efficiency over $\gamma \in [0, 1]$:

$$\tilde{\gamma} = \arg\min_{\gamma \in [0,1]} \mathrm{tr}\{J^{-1}(\beta_0 \mid \gamma)\}.$$

The Godambe information matrix may be consistently estimated by $\hat{J}\{\hat{\beta}(\gamma) \mid \gamma\} = \dot{\bar{g}}_n^{\mathrm{T}}\{\hat{\beta}(\gamma) \mid \gamma\}\bar{\Gamma}_n^{-1}\{\hat{\beta}(\gamma) \mid \gamma\}\dot{\bar{g}}_n\{\hat{\beta}(\gamma) \mid \gamma\}$. Therefore, an estimated norm is $\hat{\eta}(\gamma) = \mathrm{tr}[\hat{J}^{-1}\{\hat{\beta}(\gamma) \mid \gamma\}]$, which is the sample counterpart of the norm $\eta_0(\gamma) = \mathrm{tr}\{J^{-1}(\beta_0 \mid \gamma)\}$. The norm $\eta_0(\gamma)$ is continuous on $\gamma \in [0, 1]$ and need not be a unimodal function of γ, so there may exist multiple shrinkage coefficients that minimize $\eta_0(\gamma)$. In the implementation, greedy searching over a dense grid of γ values is desirable. Let $\gamma_0^* = \sup\{\gamma\}$ be the supremum of all such $\tilde{\gamma}$ minimizing $\eta_0(\gamma)$. The rationale for choosing the largest value of γ_0^* relates to preference of the prior dependence structure Π over the unrestricted covariance V. Although the choice of γ does not impact the result of hypothesis testing, we favour established network knowledge. In this way, a unique tuning value is obtained to achieve maximum efficiency.

The following lemma shows that the optimal shrinkage coefficient γ_0^* can be chosen consistently as the sample size goes to infinity.

LEMMA 1. *Let* $S_0 = \{\gamma : \gamma = \arg\min_{\gamma \in [0,1]} \eta_0(\gamma)\}$ *with* $\eta_0(\gamma) = \mathrm{tr}\{J^{-1}(\beta_0 \mid \gamma)\}$ *and* $S = \{\gamma : \gamma = \arg\min_{\gamma \in [0,1]} \hat{\eta}(\gamma)\}$ *with* $\hat{\eta}(\gamma) = \mathrm{tr}[\hat{J}^{-1}\{\hat{\beta}(\gamma) \mid \gamma\}]$. *Let* $\gamma_0^* = \sup\{S_0\}$ *and* $\hat{\gamma}^* = \sup\{S\}$. *Suppose* $|S_0| = |S| < \infty$ *and that both the sensitivity matrix* $G(\beta_0 \mid \gamma)$ *and the variability matrix* $\Gamma(\beta_0 \mid \gamma)$ *are bounded for* $\gamma \in [0, 1]$. *Under Conditions* A1–A6 *in the Appendix,* $\hat{\gamma}^* \to \gamma_0^*$ *in probability as* $n \to \infty$.

The proof of Lemma 1 is outlined in the Appendix. Following standard generalized-method-of-moments arguments, we establish the following theorem.

THEOREM 1. *Under Conditions* A1–A6 *in the Appendix, the regression parameter estimator* $\hat{\beta}(\hat{\gamma}^*)$ *at the optimal tuning* $\hat{\gamma}^* = \sup\{S\}$ *is asymptotically normal, i.e.,* $\sqrt{n}\{\hat{\beta}(\hat{\gamma}^*) - \beta_0\} \to N\{0, J^{-1}(\beta_0 \mid \gamma_0^*)\}$ *in distribution as* $n \to \infty$.

Theorem 1 indicates that the regression parameter estimator at the optimal shrinkage coefficient $\hat{\gamma}^*$ is asymptotically normally distributed and more efficient than other estimators obtained under an arbitrary $\gamma \in [0, 1] \setminus S$, because $\mathrm{tr}[J^{-1}\{\beta(\hat{\gamma}^*) \mid \hat{\gamma}^*\}] \leqslant \mathrm{tr}[J^{-1}\{\beta(\gamma) \mid \gamma\}]$.

4. Simulation experiment

We conducted simulations to evaluate the performance of the proposed estimator, denoted by $\hat{\beta}(\Pi^*, \hat{\gamma}^*)$, obtained under a prespecified adjacency matrix Π^* and the optimally selected shrinkage coefficient $\hat{\gamma}^*$. We consider both continuous and binary responses, and compare estimation efficiency under three different network structures: a complete network, a chain network, and a five-subregion network. Three types of correlation matrix $R(\alpha)$ are used in data generation.

N1: a complete network which uses the exchangeable correlation $R_{\mathrm{EX}}(\alpha = 0{\cdot}7)$ and $\Pi^* = M_{\mathrm{comp}}$, because M_{comp} provides an adjacency matrix of a complete network resembling a subregion of similar neuro-nodes.

N2: a chain network which uses the AR(1) correlation $R_{\mathrm{AR}}(\alpha = 0{\cdot}7)$ and $\Pi^* = M_{\mathrm{chain}}$, because M_{chain} gives an adjacency matrix of a chain network mimicking neuro-nodes along a nerve branch.

N3: two networks of five subregions with function-specific clusters specified by $R^a_{\mathrm{CL}} = \text{block-diag}\{R_{\mathrm{EX}}(\alpha = 0{\cdot}7), R_{\mathrm{AR}}(\alpha = 0{\cdot}6), I(\alpha = 0), R_{\mathrm{EX}}(\alpha = 0{\cdot}5), R_{\mathrm{AR}}(\alpha = 0{\cdot}8)\}$ and $R^b_{\mathrm{CL}} = \text{block-diag}\{R_{\mathrm{EX}}(\alpha = 0{\cdot}4), R_{\mathrm{AR}}(\alpha = 0{\cdot}6), I(\alpha = 0), R_{\mathrm{EX}}(\alpha = 0{\cdot}2), R_{\mathrm{AR}}(\alpha = 0{\cdot}8)\}$; Π^* is given by a prior target structure of the form $\Pi_{\mathrm{CL}} = \text{block-diag}\{0, M_{\mathrm{chain}}, 0, 0, M_{\mathrm{chain}}\}$.

For each scenario, 500 replications are performed, from which we obtain: the optimal shrinkage coefficient $\hat{\gamma}^*$ at each simulation using a grid search of 25 equally spaced points over $[0, 1]$; the estimation bias $(500p)^{-1}\sum_{s=1}^{500}\sum_{l=1}^{p}\|\hat{\beta}_l^{(s)} - \beta_{0l}\|$; the mean squared error $500^{-1}\sum_{s=1}^{500}\|\hat{\beta}^{(s)} - \beta_0\|_2^2$; and the total variance $500^{-1}\sum_{s=1}^{500}\mathrm{tr}\{\hat{\mathrm{var}}(\hat{\beta}^{(s)})\}$. Here $\hat{\beta}^{(s)}$ is the estimate from the sth simulation and β_0 is the true parameter. We then calculate the empirical relative efficiency and ratio of variances by calculating a ratio between the candidate and reference methods. We also examine a goodness-of-fit test and a generalized-method-of-moments-type test between nested models.

Here we present only results for continuous data; results for binary data are given in the Supplementary Material. The continuous response variables are generated from a marginal model $y_{ij} = x_{ij}^{\mathrm{T}}\beta_0 + \epsilon_{ij}$, where $x_{ij} = (x_{ij}^{(1)}, x_{ij}^{(2)})^{\mathrm{T}}$ such that $x_{ij}^{(1)}$ and $x_{ij}^{(2)}$ are generated independently from $N(j/m, 1)$ with varying means j/m over m nodes, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im})^{\mathrm{T}} \sim N\{0, R(\alpha)\}$, and $\beta_0 = (\beta_0^1, \beta_0^2)^{\mathrm{T}} = (1, 1)^{\mathrm{T}}$; n is the sample size, taken to be 50, 100 or 500, and m is the number of vertices. The sizes of the complete network N1 and the chain network N2 are set to $m = 10$ to mimic a subregion of the brain network, whereas the network of five subregions, N3, has $m = 50$, 100 or 150, with the dimension of each block set to $(m/5) \times (m/5)$.

Table 1 summarizes the biases and relative efficiencies under the three network structures, where the reference method is the oracle case, i.e., the generalized estimating equations with the true correlation, in which the correlation parameter α is set to the true value; this method is semiparametrically efficient. Here we focus on comparison of the proposed methods, including: $\hat{\beta}(\Pi = \Pi^*, \gamma = \hat{\gamma}^*)$, where both the prior structure Π^* and the unrestricted covariance V are used; $\hat{\beta}(\Pi = \Pi^*, \gamma = 1)$, where only the prior structure Π^* is used; $\hat{\beta}(\Pi = M_{\mathrm{comp}}, \gamma = 1)$, where only the prior complete network is used; $\hat{\beta}(\Pi = M_{\mathrm{chain}}, \gamma = 1)$, where only the prior

Table 1. *Empirical relative efficiency, ratio of variances, and bias of regression coefficients over* 500 *simulations, where all values have been multiplied by* 100. *The generalized estimating equations oracle is the reference method; the hybrid quadratic inference function estimator is denoted by* $\hat{\beta}(\Pi, \gamma)$ *under prior* Π *and shrinkage coefficient* γ*; empty entries represent values greater than* 10^3

		$n = 50$			$n = 100$			$n = 500$		
True network	Method	ERE	Rvar	Bias	ERE	Rvar	Bias	ERE	Rvar	Bias
Complete	$\hat{\beta}(\Pi = \Pi^*, \gamma = \hat{\gamma}^*)$	123	79	5·14	113	89	3·38	100	97	1·43
	$\hat{\beta}(\Pi = \Pi^*, \gamma = 1)$	112	91	4·91	103	96	3·27	100	99	1·42
$\Pi^* = M_{\text{comp}}$	$\hat{\beta}(\gamma = 0)$	123	79	5·14	113	89	3·38	100	97	1·43
$m = 10$	$\hat{\beta}(\Pi = M_{\text{chain}}, \gamma = 1)$	115	93	5·20	109	99	3·51	104	103	1·51
	GEE independence	118	115	5·64	116	117	3·94	120	118	1·75
	GEE unstructured			13·19	294	220	3·83	100	98	1·43
	GEE oracle($R = R_{\text{True}}$)	100	100	4·66	100	100	3·24	100	100	1·42
Chain	$\hat{\beta}(\Pi = \Pi^*, \gamma = \hat{\gamma}^*)$	113	89	4·23	107	96	2·81	101	100	1·21
	$\hat{\beta}(\Pi = \Pi^*, \gamma = 1)$	111	91	4·19	106	96	2·81	102	100	1·21
$\Pi^* = M_{\text{chain}}$	$\hat{\beta}(\gamma = 0)$	128	92	4·63	119	100	3·02	109	107	1·29
$m = 10$	$\hat{\beta}(\Pi = M_{\text{comp}}, \gamma = 1)$	137	117	4·90	138	123	3·37	131	127	1·46
	GEE independence	136	136	5·01	145	139	3·59	145	140	1·58
	GEE unstructured			13·52			6·19	101	99	1·20
	GEE oracle($R = R_{\text{True}}$)	100	100	4·00	100	100	2·72	100	100	1·19
Five-subregion	$\hat{\beta}(\Pi = \Pi^*, \gamma = \hat{\gamma}^*)$	212	86	2·41	175	107	1·66	145	127	0·64
	$\hat{\beta}(\Pi = \Pi^*, \gamma = 1)$	250	217	2·60	255	227	2·00	254	236	0·84
$\Pi^* = \Pi_{\text{CL}}$	$\hat{\beta}(\gamma = 0)$	212	114	2·44	202	141	1·80	188	166	0·73
$m = 100$	$\hat{\beta}(\Pi = M_{\text{comp}}, \gamma = 1)$	313	275	2·89	306	292	2·16	321	298	0·92
	$\hat{\beta}(\Pi = M_{\text{chain}}, \gamma = 1)$	239	189	2·52	214	202	1·77	215	209	0·77
	GEE independence	301	291	2·83	296	299	2·14	317	300	0·92
	GEE oracle($R = R_{\text{True}}$)	100	100	1·71	100	100	1·24	100	100	0·53

GEE, generalized estimating equations; ERE, empirical relative efficiency; Rvar, ratio of variances.

chain network is used; and $\hat{\beta}(\gamma = 0)$, with only the unrestricted covariance V being used. Some conventional methods are included in the comparison, namely generalized estimating equations under independence correlation, representing the independence network, under unstructured correlation, and under the true correlation. We also conducted additional simulation studies with three basis matrices in (2), but the results are not shown here due to space limitations; one of the simulations uses three basis matrices from the AR(1) correlation structure, see Fig. 2, and the other uses the three matrices I, M_{chain} and M_{comp} (Zhou & Qu, 2012). Including one more basis matrix in (2) offers little improvement in terms of empirical relative efficiency and bias. In the five-subregion network N3, results of the generalized estimating equations estimation under unstructured correlation are not provided, due to numerical failure in the case of the 100-dimensional network. In Table 1 we list results for N3 only in the case of R^a_{CL} with $m = 100$; the Supplementary Material reports full results for the other scenarios.

Table 1 shows that $\hat{\beta}(\Pi = \Pi^*, \gamma = \hat{\gamma}^*)$, the hybrid quadratic inference estimator under a prespecified adjacency matrix Π^* and the optimally selected shrinkage coefficient $\hat{\gamma}^*$, exhibits a steady fall in relative efficiency and a steady rise in the ratio of variances as n increases. It is not surprising to see that the generalized estimating equations estimator under unstructured correlation performs the worst when $n = 50$ or $n = 100$, because in this case a large number of correlations must be estimated. When the true network is the complete graph N1, the empirical relative efficiency and the ratio of variances of $\hat{\beta}(\Pi = \Pi^*, \gamma = \hat{\gamma}^*)$ are very similar to those

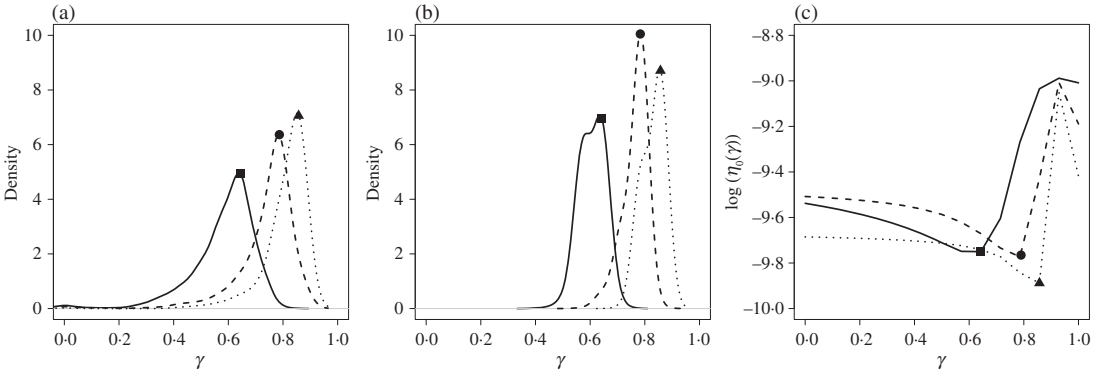

Fig. 3. Densities of $\hat{\gamma}^*$ over 500 simulations: (a) $n = 100$; (b) $n = 500$; (c) patterns of the log norm of shrinkage coefficient selection $\eta_0(\gamma) = \mathrm{tr}[J^{-1}\{\beta_0 \mid \gamma, \Pi^*, R(\alpha)\}]$ versus γ for $\hat{\beta}(\Pi = \Pi^*, \gamma = \hat{\gamma}^*)$ under the five-subregion network N3 with R^a_{CL} and $\Pi^* = \Pi_{\mathrm{CL}}$. Each panel displays the optimal γ_0^* for network size $m = 50$ (square), 100 (circle) and 500 (triangle) along with its distribution.

given by the data-driven $\hat{\beta}(\gamma = 0)$, regardless of sample size. When the true network is the chain graph N2, the performance of $\hat{\beta}(\Pi = \Pi^*, \gamma = \hat{\gamma}^*)$ becomes closer to that of the oracle generalized estimating equations as n increases. For the five-subregion graph N3, $\hat{\beta}(\Pi = \Pi^*, \gamma = \hat{\gamma}^*)$ is clearly the top performer and, in particular, is superior to $\hat{\beta}(\Pi = \Pi^*, \gamma = 1)$ and $\hat{\beta}(\gamma = 0)$.

Figure 3 displays the results of optimal shrinkage coefficient selection under R^a_{CL}, which is specified by a more realistic five-subregion network N3 with a varying network size of $m =$ 50, 100 or 150. The density plots of the selected optimal shrinkage coefficient $\hat{\gamma}^*$ for $\hat{\beta}(\Pi = \Pi^*, \gamma = \hat{\gamma}^*)$ show that the probability of $\hat{\gamma}^*$ falling near the optimal value γ_0^* increases as the sample size increases. This illustrates the selection consistency asserted in Lemma 1. Figure 3(c) shows that the target structure $\Pi^* = \Pi_{\mathrm{CL}}$ tends to receive a higher weight $\hat{\gamma}^* > 0{\cdot}5$ and hence is more informative than the unrestricted covariance V as the network size increases.

We summarize in Fig. 4 the estimation efficiency results obtained under R^a_{CL} and R^b_{CL} with $n =$ 100, 500 and $\Pi^* = \Pi_{\mathrm{CL}}$. The proposed $\hat{\beta}(\Pi = \Pi^*, \gamma = \hat{\gamma}^*)$, represented by line 2, outperforms the other approaches. When $n = 500$, the proposed method utilizing both prior and data-driven information, denoted by line 2, is clearly superior to the other approaches.

To investigate the performance of the test statistics given in § 3·2, we ran a simulation study with the following settings. The full model takes the form $y_{ij} = x_{ij}^{\mathrm{T}}\beta_0 + \theta z_i + \epsilon_{ij}$, where z_i is a subject-level variable generated from a Bernoulli distribution with probability 0·5, and x_{ij} and ϵ_{ij} are generated by the same distributions as above. The null hypothesis is $H_0 : \theta = 0$, and the alternative hypothesis is $H_1 : \theta \neq 0$. Type I error rates are computed with $\theta = 0$, while power is calculated under $\theta = 0{\cdot}2$. The size and power of the generalized-method-of-moments-type test are obtained by averaging over 25 candidate shrinkage coefficients in the range from 0 to 1 to dampen the influence of γ selection.

Table 2 summarizes the empirical Type I error and power of the test statistics at significance level 0·05 over 500 replications. The Type I error is well controlled in all cases, and the power increases as the sample size increases. Specifically, when $n = 500$, the test based on $\hat{\beta}(\Pi = \Pi^*, \gamma = 1)$ with an expert-prespecified prior target Π^* performs slightly better than the test based on $\hat{\beta}(\gamma = 0)$ for the complete or chain network. When compared with the tests based on $\hat{\beta}(\Pi = \Pi^*, \gamma \in [0, 1])$, the results are only marginally different. These results demonstrate that the null distribution for the proposed testing approach is insensitive to the choice of the prior network structure Π or of the shrinkage coefficient γ. However, the Wald test statistics, involving both $\hat{\beta}$ and $\mathrm{var}(\hat{\beta})$, depend on the selection of Π and γ.

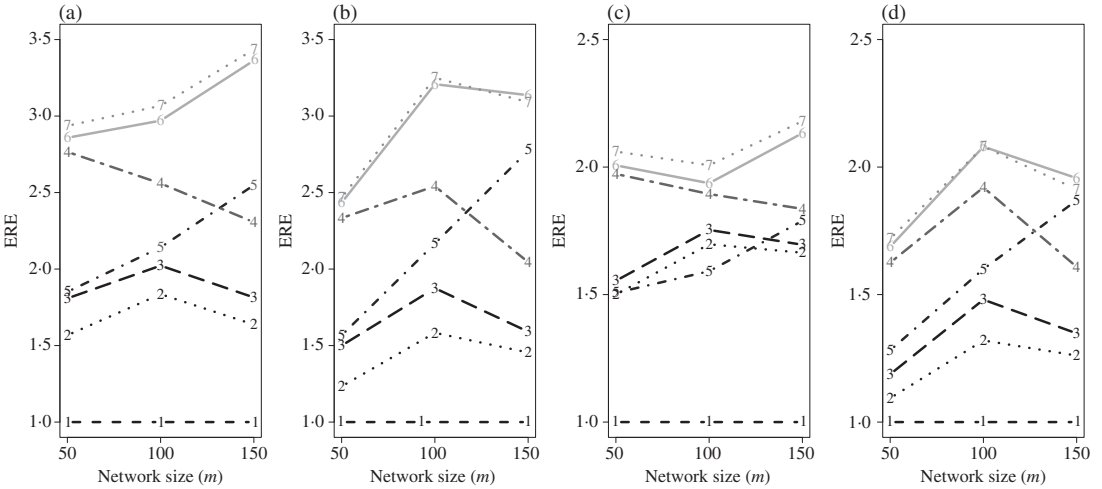

Fig. 4. Comparison of empirical relative efficiency (ERE) under the five-subregion network N3 with sample size n, number of nodes $m = 50, 100, 150$ and $\Pi^* = \Pi_{\mathrm{CL}}$: (a) R^a_{CL} and $n = 100$; (b) R^a_{CL} and $n = 500$; (c) R^b_{CL} and $n = 100$; (d) R^b_{CL} and $n = 500$. In each panel the labelled lines indicate: 1, generalized estimating equations oracle where the reference equals 1; 2, $\hat{\beta}(\Pi = \Pi^*, \gamma = \hat{\gamma}^*)$; 3, $\hat{\beta}(\gamma = 0)$; 4, $\hat{\beta}(\Pi = \Pi^*, \gamma = 1)$; 5, $\hat{\beta}(\Pi = M_{\mathrm{chain}}, \gamma = 1)$; 6, generalized estimating equations independence; 7, $\hat{\beta}(\Pi = M_{\mathrm{comp}}, \gamma = 1)$.

Table 2. *Average empirical Type I error rate and power of the test statistics* (%) *at significance level* 0·05 *over* 500 *replications; the three network structures used are the complete network N1, the chain network N2, and the five-subregion network N3 with* R^a_{CL}

		$n = 50$		$n = 100$		$n = 500$	
Network	$\hat{\beta}(\Pi^*, \gamma)$	Size	Power	Size	Power	Size	Power
Complete							
$\Pi^* = M_{\mathrm{comp}}$	$\gamma = 0$	3·0	11·4	6·2	24·2	5·2	76·4
$m = 10$	$\gamma = 1$	2·8	10·0	6·6	23·4	4·6	77·2
	$\gamma \in [0, 1]$	3·1	11·2	6·3	24·3	5·0	76·4
Chain							
$\Pi^* = M_{\mathrm{chain}}$	$\gamma = 0$	4·2	15·6	6·2	36·6	5·4	95·2
$m = 10$	$\gamma = 1$	3·8	15·4	6·0	39·6	4·6	95·8
	$\gamma \in [0, 1]$	4·0	15·4	6·1	38·0	5·1	95·5
Five-subregion							
$\Pi^* = \Pi_{\mathrm{CL}}$	$\gamma = 0$	4·4	79·6	5·6	99·2	5·6	100
$m = 100$	$\gamma = 1$	4·8	65·6	6·0	95·8	4·8	100
	$\gamma \in [0, 1]$	4·3	78·4	5·5	98·6	5·6	100

5. Data example: infant memory study

We illustrate the proposed method by applying it to the infant auditory recognition memory study discussed in § 1. Electroencephalogram data were recorded from 161 two-month-old infants using a 64-channel HydroCel Geodesic Sensor Net, from which event-related potentials were observed. Based on serum ferritin and zinc protoporphyrin levels in cord blood measured at birth, 52 of the infants were classified as iron-deficient whereas the others were classed as iron-sufficient. The primary scientific objective of this study was to evaluate the effects of prenatal and postnatal environmental exposures, such as lead and pesticides, and iron

Table 3. *Estimated regression coefficients $\hat{\beta}$ for the infant memory data with respect to mother's voice stimulus* ($*$: p-value $< 0{\cdot}05$)*, with estimated standard errors in parentheses. The first two columns of values are for $\hat{\beta}(\Pi = \Pi^*, \gamma = \hat{\gamma}^*)$ under two types of network structure suggested by our collaborators with different optimal shrinkage coefficients; the third and fourth columns of values are for $\hat{\beta}(\gamma = 0)$ and the spatial analysis-of-variance mixed-effects model; the final row lists the estimated sums of variances for $\hat{\beta}$*

Parameter	$\Pi^* = \Pi_{7\text{comp}}$ $\hat{\gamma}^* = 0{\cdot}875$	$\Pi^* = \Pi_{\text{stranger}}$ $\hat{\gamma}^* = 0{\cdot}583$	$\gamma = 0$	Spatial ANOVA mixed-effects model
age	−0·003 (0·002)	−0·003 (0·001)	−0·003 (0·001)*	−0·001 (0·002)
lead	−0·006 (0·003)	−0·005 (0·003)	−0·006 (0·003)*	0·000 (0·004)
group	0·158 (0·174)	0·158 (0·174)	0·176 (0·173)	0·587 (0·271)*
left fc	−0·803 (0·220)*	−0·854 (0·218)*	−0·811 (0·220)*	−0·824 (0·335)*
middle fc	−0·360 (0·189)	−0·338 (0·183)	−0·363 (0·186)	−0·580 (0·275)*
right fc	−1·375 (0·218)*	−1·327 (0·215)*	−1·373 (0·218)*	−1·045 (0·343)*
left po	−0·167 (0·259)	−0·359 (0·251)	−0·200 (0·251)	0·466 (0·370)
middle po	−0·056 (0·281)	−0·110 (0·280)	−0·071 (0·282)	1·566 (0·367)*
right po	0·573 (0·240)*	0·603 (0·230)*	0·559 (0·229)*	1·065 (0·392)*
group × left fc	0·714 (0·344)*	0·571 (0·380)	0·482 (0·374)	−1·143 (0·762)
group × middle fc	0·120 (0·312)	0·092 (0·339)	−0·081 (0·336)	−1·167 (0·703)
group × right fc	−0·462 (0·379)	−0·458 (0·388)	−0·612 (0·390)	−1·101 (0·746)
group × left po	0·056 (0·392)	0·167 (0·413)	0·246 (0·410)	0·207 (0·593)
group × middle po	−0·112 (0·484)	−0·080 (0·480)	−0·006 (0·491)	−0·277 (0·689)
group × right po	−1·427 (0·374)*	−1·417 (0·366)*	−1·337 (0·367)*	−0·959 (0·689)
tr{v̂ar($\hat{\beta}$)}	1·263	1·306	1·314	3·763

fc, frontal-central; po, parietal-occipital; ANOVA, analysis of variance.

deficiency on neuro-developmental outcomes. After pre-processing, the data from 56 nodes were used.

The outcome y_{ij} considered in this data analysis is a continuous variable of late slow wave activity related to the event of memory updating, which was measured as a response to the mother's voice stimulus. Nine covariates are included: centred infant age x_{i1}; centred lead concentration in cord blood x_{i2}; iron status x_{i3}, a binary measurement with 1 for iron-deficient and 0 for iron-sufficient; and six dummy variables for seven brain hemisphere regions, namely left frontal-central x_{4j}, middle frontal-central x_{5j}, right frontal-central x_{6j}, left parietal-occipital x_{7j}, middle parietal-occipital x_{8j}, right parietal-occipital x_{9j}, and other central as the reference. More details are provided in the Supplementary Material. In this analysis, interaction effects between iron status and hemisphere regions, i.e., $x_{i3}x_{4j}$, $x_{i3}x_{5j}$, $x_{i3}x_{6j}$, $x_{i3}x_{7j}$, $x_{i3}x_{8j}$ and $x_{i3}x_{9j}$, are of key interest, as they enable us to assess whether iron status could alter the amplitude of memory updating under the mother's voice stimulus over the seven brain regions. Consider the marginal linear model

$$\begin{aligned} E(y_{ij} \mid x_i) = {} & \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{4j} + \beta_5 x_{5j} + \beta_6 x_{6j} + \beta_7 x_{7j} \\ & + \beta_8 x_{8j} + \beta_9 x_{9j} + \beta_{10} x_{i3} x_{4j} + \beta_{11} x_{i3} x_{5j} + \beta_{12} x_{i3} x_{6j} + \beta_{13} x_{i3} x_{7j} \\ & + \beta_{14} x_{i3} x_{8j} + \beta_{15} x_{i3} x_{9j} \quad (i = 1, \ldots, 161;\ j = 1, \ldots, 56). \end{aligned}$$

Table 3 reports the results of regression coefficient estimation, including point estimates, standard errors and sum-of-variance estimates obtained by several methods. The methods are:

the spatial analysis-of-variance mixed-effects model, $\hat{\beta}(\gamma = 0)$, $\hat{\beta}(\Pi = \Pi^*, \gamma = 1)$, and the proposed $\hat{\beta}(\Pi = \Pi^*, \gamma = \hat{\gamma}^*)$ with the optimal tuning $\hat{\gamma}^*$. Upon consultation with our collaborators, we chose to consider two types of prior target Π^* for the hybrid quadratic inference function: one is a seven-block complete network $\Pi_{7\text{comp}} = \text{block-diag}\{M_{\text{comp}}, \ldots, M_{\text{comp}}\}$ based on the seven-block hemisphere, see Fig. 1(a), and the other is a sparse network structure learned from the separate late slow wave data under a stranger's voice stimulus using the R package space with a threshold 0·1, where the topology of Π_{stranger} is as displayed in Fig. 1(b).

As shown in Table 3, $\hat{\beta}(\Pi^* = \Pi_{7\text{comp}}, \hat{\gamma}^* = 0{\cdot}875)$ yields the smallest estimated total of variances $\text{tr}\{\hat{\text{var}}(\hat{\beta})\} = 1{\cdot}263$ of the three different hybrid quadratic inference function methods and the spatial analysis-of-variance estimator. The prior target $\Pi_{7\text{comp}}$ is favoured with $\hat{\gamma}^* = 0{\cdot}875$, and thus it is informative for unveiling the dependence of late slow wave outcomes among the 56 nodes compared to the fully data-driven covariance matrix. The second-best performer is $\hat{\beta}(\Pi^* = \Pi_{\text{stranger}}, \hat{\gamma}^* = 0{\cdot}583)$, with $\text{tr}\{\hat{\text{var}}(\hat{\beta})\} = 1{\cdot}306$, and $\hat{\gamma}^* = 0{\cdot}583$ suggests that the prior target Π_{stranger} is slightly more favourable than the data-driven dependence structure. Although these top two methods provide similar parameter estimates, the former enables us to identify more significant group-region interaction effects than does the latter. For example, the interaction effect $\hat{\beta}_{\text{group}\times\text{left fc}} = 0{\cdot}714$ is statistically significant, implying that the expected late slow wave amplitude is elevated by 0·714 units in the iron-deficient group over the iron-sufficient group in the left frontal-central subregion. Likewise, the significant interaction effect $\hat{\beta}_{\text{group}\times\text{right po}} = -1{\cdot}427$ suggests that the expected late slow wave amplitude is 1·427 units lower in the iron-deficient group than in the iron-sufficient group in the right parietal-occipital subregion. In summary, by allocating higher weights to more relevant network structures in the estimation and inference, the proposed hybrid quadratic inference function method shows promise in improving the statistical power of the networked data analysis.

6. Discussion

Although it is difficult to specify a very informative prior network topology, our simulation shows promise of improvement in efficiency when the prior structure captures part of the true network topology. That being said, our method requires estimation of a common covariance V across all subjects. In practice, networked data may not be collected from networks that have the same number of vertices and could be unbalanced due to data missingness or experimental constraints. To improve the proposed method for unbalanced networked data, the sample covariance matrix could possibly be obtained by the method of Qu et al. (2010).

Methods of sparse graph estimation are useful statistical tools for learning the target structure Π from networked data. In practice, either training data or pilot study data may not always be available. If the data are first analysed to obtain Π and then the same data reanalysed to yield results for the regression model, overfitting may occur. In such a situation, some adjustments may be needed to reach proper inference. Nevertheless, the consistency of our hybrid quadratic inference function estimation method relies only on the unbiasedness of extended scores, a feature which is independent of the choice of Π and can be justified by the goodness-of-fit test provided in the paper.

Acknowledgement

This research was supported by the U.S. National Science Foundation. We are grateful to Drs B. Lozoff and F. Geng for providing the infant memory data and scientific background on the

study. We thank the editor, associate editor and two anonymous reviewers for their constructive suggestions, which have led to a great improvement of this paper.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes additional simulation results for networked continuous data and networked binary data.

APPENDIX

The following regularity conditions are needed to establish the asymptotic properties of the hybrid quadratic inference function estimator:

Condition A1. β_0 lies in the interior of a compact parameter space $\mathcal{B} \subset \mathbb{R}^p$;

Condition A2. $g_i(\beta \mid \Pi, \gamma)$ is continuously differentiable in a neighbourhood $\mathcal{N}$ of β_0;

Condition A3. $E\{g_i(\beta \mid \Pi, \gamma)\} = 0$ for all i if and only if $\beta = \beta_0$, and $E\{\|g_i(\beta_0 \mid \Pi, \gamma)\|^2\}$ is finite, where $\|\cdot\|$ is the Euclidean norm;

Condition A4. $E\{\sup_{\beta \in \mathcal{N}} \|\partial g_i(\beta \mid \Pi, \gamma)/\partial \beta^{\mathrm{T}}\|\} < \infty$;

Condition A5. $\sqrt{n}\, \bar{g}_n(\beta_0 \mid \Pi, \gamma) \to N\{0, \Gamma(\beta_0 \mid \Pi, \gamma)\}$ in distribution as $n \to \infty$, where $\Gamma(\beta_0 \mid \Pi, \gamma) = \mathrm{cov}\{g_i(\beta_0 \mid \Pi, \gamma)\}$;

Condition A6. $J(\beta_0 \mid \Pi, \gamma) = G^{\mathrm{T}}(\beta_0 \mid \Pi, \gamma)\Gamma^{-1}(\beta_0 \mid \Pi, \gamma)G(\beta_0 \mid \Pi, \gamma)$ is nonsingular, where $G(\beta_0 \mid \Pi, \gamma) = E\{\partial g_i(\beta_0 \mid \Pi, \gamma)/\partial \beta^{\mathrm{T}}\}$.

Proof of Lemma 1

Given a target structure Π, and under the regularity conditions stated above, for a given γ, $\hat{\beta}(\gamma)$ obtained by minimizing the hybrid quadratic inference function (8) is consistent and asymptotically normal. In addition, since the weighting covariance matrix $\bar{\Gamma}_n(\hat{\beta} \mid \gamma)$ tends to $\Gamma(\beta_0 \mid \gamma)$ in probability and $\dot{\bar{g}}_n(\hat{\beta} \mid \gamma) \to G(\beta_0 \mid \gamma)$ in probability, the inverse of the Godambe information matrix $J^{-1}(\beta_0 \mid \gamma)$ of g_i may be consistently estimated by $\hat{J}^{-1}(\hat{\beta}(\gamma) \mid \gamma) = \{\dot{\bar{g}}_n^{\mathrm{T}}(\hat{\beta}(\gamma) \mid \gamma)\bar{\Gamma}_n^{-1}(\hat{\beta}(\gamma) \mid \gamma)\dot{\bar{g}}_n(\hat{\beta}(\gamma) \mid \gamma)\}^{-1}$. It follows that $\mathrm{tr}\{\hat{J}^{-1}(\hat{\beta}(\gamma) \mid \gamma)\} \to \mathrm{tr}\{J^{-1}(\beta_0 \mid \gamma)\}$ in probability as $n \to \infty$.

Write $\hat{\eta}(\gamma) = \mathrm{tr}\{\hat{J}^{-1}(\hat{\beta}(\gamma) \mid \gamma)\}$, and let $\eta_0(\gamma) = \mathrm{tr}\{J^{-1}(\beta_0 \mid \gamma)\}$. It follows that $\hat{\eta}(\gamma) - \eta_0(\gamma) \to 0$ in probability pointwise in γ on the compact set $[0, 1]$. To show that $\hat{\eta}(\gamma)$ is stochastically equicontinuous, we check a stochastic Lipschitz-type condition on $\hat{\eta}(\gamma)$: $E\{\sup_{\gamma \in [0,1]} |\partial \hat{\eta}(\gamma)/\partial \gamma|\} < \infty$. Applying Lemma 1 of Wang et al. (1986), we have

$$\frac{\partial \hat{\eta}(\gamma)}{\partial \gamma} = -\mathrm{tr}\{\hat{W}(\gamma)\hat{J}^{-2}(\gamma)\} \leqslant \rho\{\hat{W}(\gamma)\}\, \mathrm{tr}\{\hat{J}^{-2}(\gamma)\} \leqslant p\{\max_{i,j} |\hat{W}_{ij}|\}\, \mathrm{tr}\{\hat{J}^{-2}(\gamma)\},$$

where $\hat{W}(\gamma)$ is given in (A1) below and $\rho\{\hat{W}(\gamma)\}$ is the spectral radius of a $p \times p$ real symmetric matrix $\hat{W}(\gamma)$. Note that

$$\hat{W}(\gamma) = \frac{\partial \hat{J}\{\hat{\beta}(\gamma) \mid \gamma\}}{\partial \gamma} = \frac{\partial \dot{\bar{g}}_n^{\mathrm{T}}}{\partial \gamma}\bar{\Gamma}_n^{-1}\dot{\bar{g}}_n + \dot{\bar{g}}_n^{\mathrm{T}}\bar{\Gamma}_n^{-1}\frac{\partial \dot{\bar{g}}_n}{\partial \gamma} - \dot{\bar{g}}_n^{\mathrm{T}}\bar{\Gamma}_n^{-1}\frac{\partial \bar{\Gamma}_n}{\partial \gamma}\bar{\Gamma}_n^{-1}\dot{\bar{g}}_n \tag{A1}$$

with $\partial \dot{\bar{g}}_n/\partial \gamma = \partial \bar{f}_n/\partial \beta - \partial \bar{h}_n/\partial \beta$. For sufficiently large n, $\dot{\bar{g}}_n$ is continuous on $\gamma \in [0, 1]$ and hence bounded. Since $\bar{\Gamma}_n$ is also bounded and positive definite on $\gamma \in [0, 1]$, so is $\bar{\Gamma}_n^{-1}$. In addition, the expression $\bar{\Gamma}_n(\gamma) = n^{-1}\sum_{i=1}^n \{\gamma f_i + (1-\gamma)h_i\}\{\gamma f_i + (1-\gamma)h_i\}^{\mathrm{T}}$ implies that $\partial \bar{\Gamma}_n(\gamma)/\partial \gamma$ is continuous on

$[0, 1]$, and so $\partial\bar{\Gamma}_n(\gamma)/\partial\gamma$ is bounded elementwise as well. Hence, each term in (A1) is bounded elementwise on $\gamma \in [0, 1]$, and we have $\max_{i,j} |\hat{W}_{ij}| < \infty$. On the other hand, the regularity conditions ensure that $\mathrm{tr}\{\hat{J}^{-2}(\gamma)\} < \infty$. Therefore $\partial\hat{\eta}(\gamma)/\partial\gamma$ can be bounded uniformly, and the Lipschitz-type condition is satisfied. Then we obtain uniformity of convergence (Newey, 1991), $\sup_{\gamma\in[0,1]} |\hat{\eta}(\gamma) - \eta_0(\gamma)| \to 0$ in probability. Finally, since $S_0 = \{\gamma : \gamma = \arg\min_{\gamma\in[0,1]} \eta_0(\gamma)\}$ and $S = \{\gamma : \gamma = \arg\min_{\gamma\in[0,1]} \hat{\eta}(\gamma)\}$ with $|S_0| = |S| < \infty$, we have $\hat{\gamma}^* \to \gamma_0^*$ in probability as $n \to \infty$, where $\gamma_0^* = \sup\{S_0\}$ and $\hat{\gamma}^* = \sup\{S\}$.

References

Bhatia, R. (1997). *Matrix Analysis*. New York: Springer.

Fields, E. C. & Kuperberg, G. R. (2012). It's all about you: An ERP study of emotion and self-relevance in discourse. *Neuroimage* **62**, 562–74.

Gevins, A. & Smith, M. E. (2000). Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cerebral Cortex* **10**, 829–39.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–54.

Harris, D. & Mátyás, L. (1999). Introduction to the generalized method of moments estimation. In *Generalized Method of Moments Estimation*. New York: Cambridge University Press, pp. 3–30.

Hu, Y. N. & Song, P. X. K. (2012). Sample size determination for quadratic inference functions in longitudinal design with dichotomous outcomes. *Statist. Med.* **31**, 787–800.

Ledoit, O. & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Mult. Anal.* **88**, 365–411.

Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Mai, X. Q., Xu, L., Li, M. Y., Shao, J., Zhao, Z. Y., deRegnier, R. A., Nelson, C. A. & Lozoff, B. (2012). Auditory recognition memory in 2-month-old infants as assessed by event-related potentials. *Dev. Neuropsychol.* **37**, 400–14.

Meinshausen, N. & Buehlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–62.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica* **59**, 1161–7.

Newman, M. (2010). *Networks: An Introduction*. Oxford: Oxford University Press.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120–5.

Peng, J., Wang, P., Zhou, N. & Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Am. Statist. Assoc.* **104**, 735–46.

Qu, A. & Lindsay, B. G. (2003). Building adaptive estimating equations when inverse of covariance estimation is difficult. *J. R. Statist. Soc.* B **65**, 127–42.

Qu, A., Lindsay, B. G. & Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823–36.

Qu, A., Lee, J. J. & Lindsay, B. G. (2008). Model diagnostic tests for selecting informative correlation structure in correlated data. *Biometrika* **95**, 891–905.

Qu, A., Lindsay, B. G. & Lu, L. (2010). Highly efficient aggregate unbiased estimating functions approach for correlated data with missing at random. *J. Am. Statist. Assoc.* **105**, 194–204.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Siddappa, A. M., Georgieff, M. K., Wewerka, S., Worwa, C., Nelson, C. A. & DeRegnier, R. A. (2004). Iron deficiency alters auditory recognition memory in newborn infants of diabetic mothers. *Pediatric Res.* **55**, 1034–41.

Song, P. X. K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. New York: Springer.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. 3rd Berkeley Symp. Math. Statist. Prob.*, vol. 1. Berkeley: University of California Press, pp. 197–206.

Voorman, A., Shojaie, A. & Witten, D. (2014). Graph estimation with joint additive models. *Biometrika* **101**, 85–101.

Wang, S. D., Kuo, T. S. & Hsu, C. F. (1986). Trace bounds on the solution of the algebraic matrix Riccati and Lyapunov equation. *IEEE Trans. Auto. Contr.* **31**, 654–6.

Yuan, M. & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19–35.

Zhou, J. & Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *J. Am. Statist. Assoc.* **107**, 701–10.

[*Received September* 2014. *Revised November* 2015]