Bayesian Analysis of Multivariate Matched Proportions with Sparse Response

Mark J. Meyer1 · Haobo Cheng1 · Katherine Hobbs Knutson2

Received: 6 November 2022 / Revised: 24 February 2023 / Accepted: 10 March 2023 / Published online: 30 March 2023
© The Author(s) under exclusive licence to International Chinese Statistical Association 2023

Abstract
Multivariate matched proportions (MMP) data appear in a variety of contexts including post-market surveillance of adverse events in pharmaceuticals, disease classification, and agreement between care providers. It consists of multiple sets of paired binary measurements taken on the same subject. While recent work proposes methods to address the complexities of MMP data, the issue of sparse response, where no or very few “yes” responses are recorded for one or more sets, is unaddressed. The presence of sparse response sets results in the underestimation of variance components, loss of coverage, and lowered power in existing methods. Bayesian methods, which have not previously been considered for MMP data, provide a useful framework when sparse responses are present. In particular, the Bayesian probit model in combination with mean model prior specifications provides an elegant solution to the problem of variance underestimation. We examine a multivariate probit-based approach using hierarchical horseshoe-like priors along with a Bayesian functional principal component analysis (FPCA) to model the latent covariance. We show that our approach performs well on MMP data with sparse responses and outperforms existing methods. In a re-examination of a study on the system of care (SOC) framework for children with mental and behavioral disorders, we are able to provide a more complete picture of the relationships in the data. Our analysis provides additional insights into the functioning on the SOC that a previous univariate analysis missed.

Keywords Multivariate probit regression · Bayesian inference · Bayesian FPCA · Penalized Bayesian regression · Systems of care · Pediatric mental and behavioral disorders

Mark J. Meyer
mjm556@georgetown.edu

1 Department of Mathematics and Statistics, Georgetown University, Washington, DC 20057, USA
2 Duke University School of Medicine, Durham, NC 27710, USA
1 Introduction

Multivariate matched proportions (MMP) data arise when multiple sets of paired binary measurements are taken on the same subject. An example of such data come from a study of care coordination within a System of Care (SOC) framework [1]. An SOC informs the treatment of children with severe mental and behavioral disorders by coordinating care between six different components: mental health, primary care, the education system, child welfare, juvenile justice, and developmental disability services. However, primary care frequently provides the sole treatment for children with mental health disorders making it unclear if those children are receiving the potential benefits of the SOC. In particular, primary care may not be initiating contact with SOC components at the same rate as psychiatric care. To begin to investigate primary care’s tendency to initiate coordination within an SOC, Knutson et al. [1] present the results of a retrospective chart study of youth referred by pediatricians to a child psychiatrist at an urban community health center. The authors assess coordination by examining the documented contacts in the medical record between primary care and each of the remaining components of the SOC. The study then compares those contacts to the documented contacts between the psychiatrist, or specialty care, and the SOC components made after the initial psychiatric evaluation.

We summarize the contact data for all of the SOC components in Table 1, where the column headers denote the assessor, primary care or specialty care, and the column sub-headers denote the components. The original analysis of this data did not account for the multivariate nature of the data nor did it account for the sparse responses in contacts made with two components. Sparse response is most noticeable for contact made with developmental disabilities by specialty care where no contacts were made (see Table 1). The sparse response of juvenile justice requires noting that not all combinations of zeros and ones appear in the data, only 0, 0 and 0, 1. Alternatively, we can represent the data in a series of population averaged tables, as in Table 2, and diagnose sparse response by noting the zero cell counts in both developmental disabilities and juvenile justice tables. Since each component has a corresponding set of paired binary measurements, two of the five sets in this data exhibit sparse response.

Several authors consider approaches for analyzing MMP data. Klingenberg and Agresti [2] use marginal probability models to test for simultaneous marginal heterogeneity. They employ a generalized estimating equation (GEE) with an identity link and working independence to estimate effects and construct a multivariate version of McNemar’s test [3]. Some examine methods for only pairs of correlated proportions [4, 5]. Others propose multiple testing approaches to account for correlation between sets of matched proportions [6, 7]. For example, Westfall et al. [6] obtain loose control of the family wise error rate using Bonferroni–Holm on bootstrapped $p$ values. Lui and Chang [8] show that mixed-effects exponential risk models can be used to estimate the risk ratios for each set of matched proportions based on consistent estimators of the risk. In later work, the same authors consider a mixed-effects logistic regression model with
separate random effects of pair and set [9]. The authors show the estimated effect of treatment for each outcome using this approach reduces to the log of the well-known univariate Cochran–Mantel–Haenszel (CMH) estimator for stratified 2 × 2 tables [10, 11], after integrating out the random effects. Jiang and Xu [12] explore power and sample size calculations for this and the GEE-based model. As is common in the analysis of matched proportions, each of these methods exclude the concordant pairs, relying only on information from the discordant pairs [13].

Of note is that two of the existing methods for MMP data cannot directly handle sparse responses, as we will now illustrate. The GEE-based approach estimates marginal probabilities using the identity link under the assumption

Table 1 Multivariate matched proportions data from the SOC study for care-specific contact with developmental disabilities (DD), mental health (MH), juvenile justice (JJ), child welfare (CW), and education system (ED)

|                  | Primary care | Specialty care | Pattern count |
|------------------|--------------|----------------|---------------|
|                  | DD MH JJ CW ED | DD MH JJ CW ED |               |
| 0 0 0 0 0 0       | 0 0 0 0 0 1   | 17             |
| 0 0 0 0 0 0       | 0 0 0 0 0 0   | 16             |
| 0 1 0 0 0 0       | 0 1 0 0 0 0   | 5              |
| 0 0 0 0 1 0       | 0 0 0 0 0 1   | 5              |
| 0 1 0 0 0 0       | 0 1 0 0 0 1   | 3              |
| 0 0 0 0 0 0       | 0 1 0 0 0 0   | 3              |
| 0 0 0 0 0 0       | 0 1 0 0 0 1   | 3              |
| 0 1 0 1 0 0       | 0 0 0 0 0 0   | 2              |
| 0 0 0 0 0 1       | 0 0 0 0 0 0   | 2              |
| 0 0 0 0 0 0       | 0 0 0 0 0 1   | 2              |
| 1 0 0 0 0 0       | 0 0 0 0 0 0   | 1              |
| 0 1 0 0 0 0       | 0 0 0 0 0 0   | 1              |
| 0 1 0 1 0 0       | 0 0 0 0 0 1   | 1              |
| 0 0 0 0 0 1       | 0 0 0 0 0 1   | 1              |
| 0 1 0 0 1 1       | 0 0 0 0 1 1   | 1              |
| 0 0 0 0 0 1       | 0 0 0 0 1 1   | 1              |
| 0 1 0 0 0 0       | 0 0 0 0 0 0   | 1              |
| 0 0 0 0 0 0       | 0 0 0 0 0 0   | 1              |
| 0 0 0 1 0 0       | 0 0 0 1 0 1   | 1              |
| 0 0 0 0 0 1       | 0 0 1 0 0 1   | 1              |
| 0 1 0 1 0 0       | 0 0 0 0 0 1   | 1              |
| 0 0 0 1 0 0       | 0 0 0 1 1 1   | 1              |

Table values are 1 if contact was made with the agency in the column sub-header by the assessor in the column header, and 0 if no contact was made. The Pattern Count column denotes the number of times the pattern appears in the data. This table excludes patterns that do not appear in the data.
Table 2  Population averaged tables for all SOC components

| Development disabilities | Mental health |
|--------------------------|---------------|
| | Specialty care | Primary care |
| Specialty care | Yes | No |
| Yes | $n_{111} = 0$ | $n_{121} = 0$ |
| No | $n_{211} = 3$ | $n_{221} = 71$ |
| | Yes | No |
| No | $n_{112} = 12$ | $n_{122} = 12$ |
| | $n_{212} = 4$ | $n_{222} = 46$ |
| | Yes | No |
| Child welfare | Specialty care | Primary care |
| Specialty care | Yes | No |
| Yes | $n_{114} = 3$ | $n_{124} = 7$ |
| No | $n_{214} = 4$ | $n_{224} = 60$ |
| | Yes | No |
| No | $n_{115} = 11$ | $n_{125} = 31$ |
| | $n_{215} = 2$ | $n_{225} = 30$ |
| | Yes | No |
| Juvenile justice | Specialty care | Primary care |
| Specialty care | Yes | No |
| Yes | $n_{113} = 1$ | $n_{123} = 2$ |
| No | $n_{213} = 0$ | $n_{223} = 71$ |

‘Yes’ denotes contact was made with the component, ‘No’ indicates contact was not made. Within each table, the diagonal counts within each table represent the number of concordant pairs while the off-diagonal counts represent the number of discordant pairs.
of working independence [2]. Using a model with one parameter for each marginal probability, the authors show that the estimates equal the sample probabilities of the columns of Table 1. But when a column is sparse, as is the case for the developmental disabilities component of the SOC data, the estimated probability is zero. The estimated variance will also be zero, as will the associated covariances. The variance of the difference in marginal probabilities between care type depends on both this variance and the corresponding covariance of the sparse column, resulting in underestimation of the variance of the difference. The bootstrap-based method pre-differences the paired outcomes, i.e., takes the difference between the paired columns of Table 1, and estimates the differences in the marginal probabilities using sample averages [6]. In general, these differences can take on the values $-1, 0, \text{ and } 1$. But when the response is sparse, as in the developmental disabilities component, the differences can only be 0 or 1 meaning the lower bound of a bootstrapped interval is bounded below by 0. Similarly, for the juvenile justice component, only $-1$ and 0 are possible realizations from the data resulting in a bootstrapped interval that is bounded above by 0. Once again, the uncertainty of the difference in marginal probabilities for sparse outcomes is underestimated.

Lui and Chang do both provide adjustments for sparse data in both their exponential risk model (ERM) and their mixed-effects regression approach [8, 9]. The ERM approach, for example, bases its estimation off of the $K$-specific tables in Table 2. Specifically, the estimate is the ratio of the counts of discordant pairs in each table: $n_{21k}/n_{12k}$. When one of these elements is zero, the authors suggest adding $1/2$ to the numerator and the denominator making their sparsity adjusted estimator $(n_{21k} + \frac{1}{2})/(n_{12k} + \frac{1}{2})$. Lui and Chang do not evaluate the performance of their estimators in the presence of sparse response and instead evaluate the performance of their testing procedure in non-sparse simulated data settings [8, 9]. Such a correction is philosophically Bayesian in its thinking as it recognizes, a priori, that the true population parameter is likely not zero (or infinite) and adjusts the estimate accordingly. In fact, this estimator is an empirical Bayes estimate of the marginal risk for a single matched set under Jeffreys’ prior when using the Dirichlet model for univariate matched proportions proposed by Altham [14].

Adjustments of this kind are not generalizable either: it is not applicable to the GEE or Bootstrap approaches as it adjusts the counts from Table 2, whereas both the GEE and Bootstrap rely on the individual-level data as in Table 1. Other common adjustments like data augmentation, e.g., adding one success and one failure in the case of binary data, do not easily generalize to the multivariate case either. Augmenting sparse cells in Table 2 separately ignores the multivariate nature of the data, breaking potential underlying correlations. Adding covariate patterns to Table 1 must take into account the impact on the multivariate structure and can lead to severe sample size inflation. At the extreme, augmentation would involve adding one of every possible type of covariate pattern to Table 1. For even a small number of sets, such an addition would greatly inflate the sample size far beyond what augmentation strategies aim for, potentially biasing the results. Thus, there is a need for an MMP method that can accommodate sparse responses more adequately.
A Bayesian approach can remedy the underestimation of the variance estimation in the presence of sparse response via model and prior specification. The literature on Bayesian methods for matched proportions is limited to the univariate case. For example, Altham [14] examines closed form results for a multinomial model with Dirichlet priors while Broemeling and Gregurich [15] discuss the Gibbs sampling approach to this model. Ghosh et al. [16] present an item response model with logit, probit, and complementary log–log links. This formulation allows for a hierarchical Bayesian modeling of the data. The authors show that, for such models, the concordant pairs also contribute to estimation and inference. While these approaches are useful for univariate matched proportions, to our knowledge, no previous work considers Bayesian methods for the analysis of MMP data.

Using the Albert and Chib [17, 18] latent variable representation of the probit regression model as a basis, we propose a Bayesian approach for multivariate matched proportions. There are two sources of variability in the latent probit model: the errors on the latent variable itself and the prior variance on the model components. The first source is addressed immediately by using the Albert and Chib model: because the transformation is scale invariant, the variance on the latent errors is set to one. The second source of variation, via the prior on the model components, requires estimating the variance which, in turn, shrinks the coefficients toward the prior mean via the Bayesian version of the well-known mixed model formulation of penalized regression—see Chapter 4 of Ruppert, Wand, and Carroll’s *Semiparametric Regression* [19] for an overview of the relationship between mixed models and penalized regression. We consider a global hierarchical prior using half-Cauchy priors on the variance components [20] in a prior that is similar to the horseshoe prior [21]. To account for the multivariate nature of the data, our approach makes use of a Bayesian functional principal components analysis or FPCA [22] to model a general latent covariance matrix. The Bayesian FPCA can accommodate a wide range of covariance structures, not just those arising from functional data, making it an appealing approach for dealing MMP data. In simulation, we demonstrate that our approach has better coverage and power in the presence of sparse responses than the existing methods for MMP data while maintaining similar absolute bias and interval widths under a variety of settings. Finally, we reanalyze the SOC data using our approach and include the two components with sparse responses which were excluded from the original analysis. Our reanalysis confirms some of the original findings but identifies additional results the univariate analysis missed.

The remainder of the manuscript is organized as follows: Sect. 2 details our multivariate probit-based approach. Section 3 presents the findings of our simulation study. In Sect. 4, we describe our reanalysis of the SOC data. Finally, in Sect. 5, we provide a discussion of our approach.

2 Methods

At the pair-level, let $x_{ijk}$ denote the response for pair $i$, at observation $j$, for the $k$th set where $i = 1, \ldots, n$, $j = 1, 2$, and $k = 1, \ldots, K$. In context of the SOC data, $i$ denotes the subjects and $j$ denotes care type, primary care or specialty care. The
index \( k \) denotes the SOC component with which contact is made, \( x_{ijk} = 1 \), or not made, \( x_{ijk} = 0 \). Combining across \( k \), we let \( x_j = [x_{ij1} \ldots x_{ijk}]' \) be the vector of responses at observation \( j \) for pair \( i \). Subject \( i \)'s stacked vector of responses is then \( x_i = [x_{i1} \ x_{i2}]' \). Realizations of this vector make up the rows of Table 1.

The target of inference is a vector consisting of the differences in the \( k \)-specific marginal probabilities of contact, \( P(x_{ijk} = 1) = \theta_{jk} \), which both the GEE-based [2] and bootstrap-based [6] methods estimate. First let \( \theta = [\theta_{11} \ldots \theta_{1K} \ \theta_{21} \ldots \theta_{2K}]' \) be the vector of marginal probabilities. Then let \( \rho_k \) be the \( k \)th difference in the marginal probabilities, thus \( \rho_k = \theta_{1k} - \theta_{2k} \). The vector of \( \rho_k \)'s is \( \rho = [\rho_1 \ldots \rho_K]' \). Alternatively, define \( \mathbf{L} \) to be a block matrix of the form \( \mathbf{L} = (\mathbf{I}_k - \mathbf{I}_k) \), where \( \mathbf{I}_K \) is a \( K \times K \) identity matrix. The vector representing the differences in marginal probabilities is then \( \rho = \mathbf{L}\theta \).

Our approach begins with the Bayesian probit regression as first described by Albert and Chib [17, 18]. The regression framework makes use of the latent variable representation of the probit model. In the Bayesian context, this approach can be considered a data augmentation step where the latent variables are the augmented data [23]. We define the following mapping:

\[
x_{ijk} = \begin{cases} 
0, & \text{if } z_{ijk} < 0 \\
1, & \text{if } z_{ijk} \geq 0
\end{cases}
\]  

(1)

where the latent model might be \( z_{ijk} = \beta_{jk} + \epsilon_{ijk} \). In general, \( \epsilon_{ijk} \sim N(0, \sigma^2) \), but because this latent variable transformation is scale invariant, \( \sigma^2 \) is not identifiable and is typically set to 1 [17]. The model is saturated, so the relationship between the latent space coefficients, \( \beta_{jk} \), and the marginal probabilities, \( \theta_{jk} \), is \( \beta_{jk} = \Phi^{-1}(\theta_{jk}) \), where \( \Phi(\cdot) \) is the standard normal CDF. This relationship, and indeed the probit model itself, is induced by assuming the error terms are normal.

The model in Eq. 1 assumes independence in the latent space which, if applied to MMP data, would imply independence in the data itself. To account for possible correlations in the multivariate sample, the \( z_{ijk} \) for a fixed subject \( i \) are combined into a latent vector, specifically \( \mathbf{z}_i = [z_{i11} \ldots z_{i1K} \ z_{i21} \ldots z_{i2K}]' \). The latent model is then \( \mathbf{z}_i = \mathbf{\beta} + \mathbf{\epsilon}_i \), where \( \mathbf{\beta} = [\beta_{11} \ldots \beta_{1K} \ \beta_{21} \ldots \beta_{2K}]' \) and \( \mathbf{\epsilon}_i = [\epsilon_{i11} \ldots \epsilon_{i1K} \ \epsilon_{i21} \ldots \epsilon_{i2K}]' \). The latent error vector, \( \mathbf{\epsilon}_i \), is then assumed to be multivariate normal, \( \mathbf{\epsilon}_i \sim N(\mathbf{0}, \mathbf{\Sigma}) \) for the correlation matrix \( \mathbf{\Sigma} \). Chib and Greenberg [24] discuss this multivariate extension of the Albert and Chib [17, 18] model.

Because of its structure, drawing posterior samples for \( \mathbf{\Sigma} \) can be difficult. Chib and Greenberg [24] suggest a Metropolis–Hastings step to obtain draws from the posterior of \( \mathbf{\Sigma} \) and others authors have further discussed the same issue [25–27]. In all of these approaches, the MCMC sampler also requires obtaining draws from a truncated multivariate normal posterior, the posterior conditional distribution of \( \mathbf{z}_i \), for each subject. Using existing approaches for randomly sampling from truncated multivariate normal distributions is time intensive. In our preliminary model exploration, a model of this kind took over 11 h to generate 20,000 posterior samples.
when \( K = 2 \) (run on a computer with a 3.7 GHz 6-Core Intel Core i5 processor and 32 GB of memory).

To accommodate a multivariate structure without inducing such a computational burden, we turn to recent developments in the functional data literature. In particular, Goldsmith and Kitago [28] and Meyer et al. [29] both show that the Bayesian FPCA [22] can be used to model a variety of covariance structures for both Gaussian [28] and ordinal [29] functional data including independent, exponential, and compound symmetric (also known as exchangeable) covariance structures. For MMP data, we may expect exchangeability within observation level \( j \), what might be called a block compound symmetric structure, or independence within \( j \) but dependence between \( k \). A more general form or unstructured form may result as well. An additional challenge is that it is hard to know what the structure of the latent correlation matrix might be. Thus, we propose using an FPCA in the latent space to model possible multivariate associations in MMP data.

We perform the FPCA on the latent errors setting \( e_i = \Omega \psi c'_i + \epsilon_i \) for the \( M \times 1 \) vector of subject scores \( c_i \), the \( 2K \times M \) matrix of basis coefficients \( \Psi \), and the \( 2K \times 2K \) matrix of basis functions \( \Omega \), where \( M \) is number of principal components used in the decomposition. The model for each subject’s latent vector is then

\[
z_i = \beta + \Omega \psi c'_i + \epsilon_i,
\]

where \( \epsilon_i \) is a \( 2K \times 1 \) vector of normal, independent errors with variance equal to \( \sigma^2 \Sigma 2K \). Consistent with the previous work using Bayesian FPCA, we use \( M = 2 \) scores for the expansion \([28, 29]\). Letting \( \Omega \) be a matrix of cubic B-spline basis functions, the prior for \( \Psi \) is matrix normal \([30]\). A matrix normal prior corresponds to placing a multivariate normal prior on the vectorized version of the matrix. Let \( \text{vec}(-) \) be a function that vectorizes matrices. The prior is then \( \text{vec}(\Psi) \sim N(0, (\Lambda_{\psi} \otimes P)^{-1}) \), where \( \Lambda_{\psi} \) is an \( M \times M \) diagonal matrix of tuning parameters \( \lambda_m^{-1} \) for \( m = 1, \ldots, M \) and \( P \) is a \( 2K \times 2K \) penalty matrix of the form \( P = \xi P_0 + (1 - \xi)P_2 \). The matrices \( P_0 \) and \( P_2 \) correspond to zeroth and second derivative penalty matrices for B-splines, respectively—see Eilers and Marx’s work on penalized B-splines \([31]\), or \( p \)-splines, for details. The value \( \xi \) is a control parameter that balances smoothness and shrinkage with values near zero favoring shrinkage—we take a value consistent with prior work using Bayesian FPCAs, \( \xi = 0.01 \) \([28, 29]\). The subject scores, \( c_i \), have mean zero normal priors with variance equal to \( I_M \), an \( M \times M \) identity matrix. For the variance component, we place a non-informative prior on \( \sigma^2, \sigma^2 \sim IG(1, 1) \), which will be updated only for the purpose of improving model fit on \( \beta \). The tuning parameter priors are taken to be non-informative as well, thus \( \lambda_m \sim IG(1, 1) \).

We now consider the prior on the components of \( \beta \). To accommodate sparse responses, we must regularize these parameters. There are many choices for shrinkage priors in the Bayesian framework, but we select normal priors with a half-Cauchy hyper-prior on the scale. Gelman \([20]\) proposed the use of half-t priors for variance components, of which the half-Cauchy is a special case. Polson and Scott \([32]\) show that the half-Cauchy performs well as global-shrinkage prior and recommend their routine use in normal hierarchical models. Noting that
\( \beta = [\beta_{11} \cdots \beta_{1K} \beta_{21} \cdots \beta_{2K}] \), we place the following global hierarchical prior on the \( \beta_{jk} \):

\[
\beta_{jk} \sim \mathcal{N}(0, \lambda)
\]

\[
\lambda \sim \text{IG} \left( \frac{1}{2}, \frac{1}{\mu} \right)
\]

\[
\mu \sim \text{IG} \left( \frac{1}{2}, \frac{1}{A^2} \right)
\]

for \( i = 1, 2 \), and \( j = 1, \ldots, J \) and, where \( \text{IG} \) denotes the inverse gamma distribution and \( A \) is a fixed hyper-parameter. These hierarchical inverse gamma priors are the mixture representation of the half-Cauchy prior [33]. That is, if \( \lambda \sim \text{IG}(1/2, 1/\mu) \) and \( \mu \sim \text{IG}(1/2, 1/A^2) \), then \( \sqrt{\lambda} \sim \text{HC}(A) \), where \( \text{HC} \) denotes the half-Cauchy. The choice of the hyper-parameter \( A \) can be regarding as placing a weak upper bound on the variance of the \( \beta \) [20]. The parameter can be interpreted in context as the standard deviation of the corresponding model component [20]. Since we are modeling in a latent space, our choice of the value of \( A \) is less clear as the scale is not directly dependent on a covariate. Thus, we select a relatively large value, \( A = 10 \). This prior is also similar to the popular horseshoe prior for regularizing model coefficients [21].

The mixture-model representation of the half-Cauchy induces a model with fully identifiable conditional posterior distributions. Combining with the FPCA specification, we can obtain the full conditional posterior distributions for each model component. Let \( \Psi_{ijk} \) denote the \( j, k \)th sub-components of the vector \( \mathbf{c}' \mathbf{Z} \) be the vector of all \( z_{ijk} \)'s, and \( \mathbf{W} \) be a design matrix built by stacking \( 2K \times 2K \) identity matrices on top of each other. The full conditionals are given by

\[
z_{ijk} | x_{ijk} = 0, \beta_{jk}, \Psi_{ijk} \sim \mathcal{N}(\beta_{jk} + \Psi_{ijk}, 1)_{\{z_{ijk} < 0\}}
\]

\[
z_{ijk} | x_{ijk} = 1, \beta_{jk}, \Psi_{ijk} \sim \mathcal{N}(\beta_{jk} + \Psi_{ijk}, 1)_{\{z_{ijk} \geq 0\}}
\]

\[
\beta | Z, W \sim \mathcal{N}\left[ \Sigma_{\beta} W' (Z - \Psi), \Sigma_{\beta} \right]
\]

\[
\lambda | \beta, \mu \sim \text{IG} \left( K + \frac{1}{2}, \frac{1}{\mu} + \frac{1}{2} \beta' \beta \right)
\]

\[
\mu | \lambda \sim \text{IG} \left( 1, \frac{1}{A^2} + \frac{1}{\lambda} \right)
\]

\[
\text{vec}(\Psi)_{\text{rest}} \sim \mathcal{N} \left[ \frac{1}{\sigma^2_{\epsilon}} \Sigma_{\Psi} (C \otimes \Omega)'(Z - W\beta), \Sigma_{\Psi} \right]
\]

\[
c_{i} | \text{rest} \sim \mathcal{N} \left[ \frac{1}{\sigma^2_{\epsilon}} \Sigma_{c} (\Omega \Psi)'(z_{i} - \beta), \Sigma_{c} \right]
\]

\[
\sigma^2_{\epsilon} | \text{rest} \sim \text{IG} \left[ 1 + nK, 1 + \frac{1}{2} (Z - W\beta - \Psi)'(Z - W\beta - \Psi) \right]
\]

\[
\lambda_{m} | \text{rest} \sim \text{IG} \left( 1 + M/2, 1 + \frac{1}{2} \Psi_{m}' \mathbf{P} \Psi_{m} \right)
\]
\[ \Sigma_p = \left( \frac{1}{\sigma_p^2} W'W + \lambda^{-1} I_{2K} \right)^{-1}, \quad \Sigma_\psi = \left( \frac{1}{\sigma_\psi^2} C'C \otimes \Omega'\Omega + \Lambda_\psi \otimes P \right)^{-1}, \]

\[ \Sigma_c = \left[ \frac{1}{\sigma_c^2} \Omega \Psi' (\Psi \Omega) + I_M \right]^{-1}, \]

where \( C \) equal to the \( n \times M \) matrix formed by stacking the \( c_i \) vectors on top of each other, \( \Psi_m \) equal to the \( m \)th column of \( \Psi \), and \( \psi \) the vector containing all \( \Psi_{ijk} \)s. The subscript notation \( \{ \cdot \} \) on the normal densities indicates truncation to the region defined in the brackets. The function \( \text{vec}(\cdot) \) vectorizes the matrix \( \Psi \), making it a \( 2KM \times 1 \) vector. Note that we draw samples from \( \sigma_\epsilon^2 \) to improve the estimation of other model components, but fix \( \sigma_\epsilon^2 \) at one when sampling the latent variable for identifiability.

Using the FPCA for the latent correlation does not require sampling from a multivariate truncated normal nor does it necessitate a Metropolis–Hastings step to draw latent correlations. Estimates of \( \rho \) are easily obtained from the posterior of \( \beta \) using \( \rho = L\Phi(\beta) \) with \( L \) as defined above. This model is much more computationally efficient than other multivariate probit models taking between 60 and 70 s, depending on \( K \), to generate 20,000 posterior samples (on a computer with a 3.7 GHz 6-Core Intel Core i5 processor and 32 GB of memory). For ease of implementation, we have made our code publicly available at https://github.com/markjmeyer/BMMP.

### 3 Simulation Study

We base our simulation study off of the sample size and effect sizes observed in the SOC data. For each dataset, we generate data for \( n = 75 \) subjects, where \( K = 2, 3, 4, \) and \( 5 \) sets of binary measurements are taken on each subject. We aim to induce sparse responses within sets and achieve this by first generating probabilities for \( \theta \) from a multivariate normal distribution with varying correlation structures. In the event a draw from this distribution is negative, it is set to zero. Values of \( x_i \) are then drawn from a Bernoulli distribution using the potentially correlated \( \theta \). The “true” values of \( \theta \) vary with \( K \):

\[
\begin{align*}
K = 2 : \theta &= \begin{bmatrix} 0.05 & \theta_{12} & 0.25 & 0.005 \end{bmatrix}, \\
K = 3 : \theta &= \begin{bmatrix} 0.05 & \theta_{12} & 0.1 & 0.25 & 0.005 & 0.15 \end{bmatrix}, \\
K = 4 : \theta &= \begin{bmatrix} 0.05 & \theta_{12} & 0.005 & 0.1 & 0.25 & 0.005 & 0.05 & 0.15 \end{bmatrix}, \text{ and} \\
K = 5 : \theta &= \begin{bmatrix} 0.05 & \theta_{12} & 0.005 & 0.1 & 0.25 & 0.25 & 0.005 & 0.05 & 0.15 & 0.35 \end{bmatrix}.
\end{align*}
\]

The \( k = 2 \) component is set to exhibit sparse response since \( \theta_{22} = 0.005 \). We add an additional sparse responses when \( K = 4 \) and \( 5 \), although sparse responses may occur for any component in any given simulated draw. To evaluate the effect of sparse response as the degree of difference between observations changes, we vary \( \theta_{12} \) from 0.05 to 0.2 by 0.1. All of these values for \( \theta \) encompass the range of values we see when estimating \( \rho \) for the SOC data. The underlying covariance structure in the simulation is either independent or block compound symmetric with correlation equal to 0.5. For example, under the block compound symmetry scenario, the \( K = 2 \) and \( K = 3 \) covariances take the forms
respectively. We select this structure to represent possible within care-type correlation while assuming that between care-type correlation is zero, since they are different evaluators.

To evaluate the MMP methods, we determine the coverage, power, absolute bias, and interval width for the $k = 2$ component, $\rho_2 = \theta_{21} - \theta_{22}$, which is simulated to exhibit sparse response. We compare our multivariate probit analysis approach from Sect. 2 to the GEE-based approach [2], the Bootstrap-based method [6], and a modified ERM approach [8]. The ERM estimates risk ratios [8], however, for comparison with the GEE, Bootstrap, and the Bayesian approaches, we estimate the risk difference using their model. All Bayesian estimates are based on 20,000 total samples, discarding the first 10,000 while all Bootstrap estimates are based on 10,000 resamples. Additional details on the GEE, ERM, and Bootstrap approaches are in Sect. 1 of the Supplementary Materials.

For each combination of $K$, $\theta_{12}$, and correlation structure, we generate 1000 simulated datasets, retaining only those datasets that exhibit sparse response for the second set. Sparse response is determined as described in Sect. 1 and as exhibited in either Tables 1 or 2. Thus, if either column corresponding to $k = 2$ is all zeros or if the population average table for the $k = 2$ set contains zeroes, then the set is considered to have a sparse response. The parameters of the simulation result in between 285 and 422 datasets exhibiting sparse responses (out of 1000), depending on $\theta_{12}$, $K$, and the correlation structure. We then randomly sample 200 of these datasets, under each combination, for evaluation. Results were similar between correlation structures although the models perform slightly better when the true structure is independent. We only present results from the block compound symmetric simulation here with results from the independent structure simulation in Sect. 2 of the Supplementary Materials. Code to run all of our simulations is available at https://github.com/markjmeyer/BMMP.

Figure 1 contains the empirical coverage for all values of $\theta_{12}$ and $K$ under a “true” block compound symmetric correlation structure. All intervals were constructed at the 95% level. Coverage curves for the Bayesian multivariate probit analysis (MVP) are in solid gold with circles, the exponential risk model (ERM) results are in dashed red with triangles, the GEE-based model (GEE) results are in dotted green with plus signs, and the Bootstrap-based model (Boot) results are in dotted-dashed blue with x’s. Consistent with other evaluations of intervals for binary data [34, 35], we find
that the coverage for all methods varies by effect size. The MVP has the highest coverage for almost every combination of $\theta_{12}$ and $K$, is typically at or above nominal (95%), and never drops below 91%. The comparator methods, however, have coverage that is frequently below nominal and, in some cases, is at or below 90%. Coverage is more inconsistent for the ERM and GEE, particularly when $\theta_{12}$ is smaller. As $\theta_{12}$ increases, the coverage for all methods nears nominal. Coverage is slightly higher for all methods when the true correlation structure is independent.

A gain in coverage can come at a loss in power, so we also investigate the power of each method to test the null $H_0 : \rho_2 = 0$. The results are in Fig. 2 where significance is determined by exclusion from the 95% interval for the MVP and Bootstrap-based model and by the resulting $p$ values being less than 0.05 for the GEE and ERM. Power curves for the models follow the same color, line, and point schemes as in the depiction of the coverage curves. The MVP has the highest power at each $\theta_{12}$,
regardless of the value of $K$. The GEE and Bootstrap are second with nearly identical power while the ERM has the lowest power. All models have power that goes to one as $\theta_{12}$ increases. Power for all models is slightly higher across $K$ when the true underlying correlation structure is independent, see Fig. 2 of the Supplementary Materials.

Finally, we assess the absolute bias of each method, in Fig. 3, and interval width, in Fig. 4. These figures contain boxplots aggregating absolute bias and width across all $\theta_{12}$ for each $K$. The same color schemes used in the coverage and power curves are used for the absolute bias and interval width boxplots. Absolute bias is virtually identically between the four methods. Each boxplot in Fig. 3 has similar quartiles and overall shape. Absolute bias also remains consistent as $K$ increases, although when $K = 2$ there are more simulated datasets with outlying absolute bias values. Interval width in Fig. 4 is very similar between the four methods and the range of
widths is about the same no matter the value of \( K \). The medians differ slightly with the MVP tending to have the widest intervals followed by the ERM, the GEE-based model, and then the Bootstrap-based method. Changes in \( K \) do not appear to have much of an effect on width. Absolute bias and interval width behave similarly when the true correlation structure is independent.

4 Reanalysis of the SOC Data

The goal of the SOC study is to determine if primary care physicians are initiating contact with the same agencies within the SOC framework, and with similar frequency, to psychiatric, or specialty, care. Knutson et al. argue that it is important to identify services for improved coordination to help the overburdened mental health system and their original findings suggest there is room for improvement.
among primary care givers [1]. They separately analyze only the non-sparse sets using standard univariate methods including logistic regression [1]. Because of the sparse response in the juvenile justice and developmental disabilities components, the authors do not examine these outcomes. We now jointly reanalyze the data using all outcomes, presenting the results of our model here and the results using the GEE and Bootstrap-based model in Sect. 3 of the Supplementary Materials.

In total, we have 74 patients available for analysis. The most common primary mental health diagnosis was mood disorder followed by attention-deficit hyperactivity disorder, adjustment disorder, and anxiety disorder [1]. Additional primary diagnoses include post-traumatic stress disorder, substance abuse, and learning disorder [1]. Each patient was initially seen by primary care, who made recommendations for contact (or not) with components of the SOC. Separately, specialty care later examined each patient and independently made recommendations
for contact within the SOC. We only seek to estimate and analyze $\rho$, the differences in the marginal probability of contact between primary care and specialty care for each SOC component. Table 3 presents the results of the analysis using the MVP approach including posterior median values for each $\rho_k$, 95% credible intervals (CrI), and posterior probabilities that $\rho_k$ is greater than zero, $P(\rho_k > 0)$. In Table 3, we abbreviate each component as follows: developmental disabilities (DD), mental health (MH), juvenile justice (JJ), child welfare (CW), and education system (ED). In total, we generated 20,000 draws from the posterior, retaining the last 10,000 for estimation and inference. All model parameters were judged to have converged using trace plots (see Sect. 3 of the Supplementary Materials) and the Gelman–Rubin diagnostic [36] (see Table 3, where values of $\hat{R}$ and its upper 95% CrI near one suggest convergence). As a sensitivity analysis, we vary the choice of $A$ in the hierarchical prior considering $A = 1, 5, 10$. We find no qualitative difference in the results by varying these priors and only minor quantitative differences. Consequently, we present the results for $A = 10$. Tables similar to Table 3 when $A = 1$ and 5 are in the Supplementary Materials.

We begin by noting what is consistent in our analysis with the univariate analysis from Knutson et al. [1]. The authors found a significant association for the education component of the SOC whereby primary care contacted the education system less than specialty care. We confirm this result noting that the difference in the marginal probability of contact with the education system between primary care and specialty care is $\rho_{ED} = -0.394$ (95% CrI $[-0.535, -0.242]$, $P(\rho_{ED} > 0) = 0.000$). Knutson et al. [1] also found no significant associations for the child welfare and mental health components. Our analysis confirms this result for contact with child welfare ($\rho_{CW} = -0.036$, 95% CrI $[-0.144, 0.068]$, $P(\rho_{CW} > 0) = 0.242$) but by analyzing the data using the MVP, we have a more nuanced view of the relationship between care type and contact with mental health. Specifically, the difference in contact with mental health between primary care and specialty care is $\rho_{MH} = -0.109$ and its 95% CrI includes zero ($-0.254, 0.037$). But only 7% of the posterior distribution of $\rho_{MH}$ sits above zero which indicates that a large portion of the posterior distribution is negative.

| Component | $\rho$ | Cred. Int. | $P(\rho > 0)$ | Gelman–Rubin |
|-----------|--------|------------|---------------|--------------|
| DD        | 0.036  | -0.001     | 0.095         | 0.972        | 1.003 1.008 |
| MH        | -0.109 | -0.254     | 0.037         | 0.070        | 1.001 1.003 |
| JJ        | -0.021 | -0.081     | 0.028         | 0.165        | 1.003 1.007 |
| CW        | -0.036 | -0.144     | 0.068         | 0.242        | 1.001 1.003 |
| ED        | -0.394 | -0.535     | -0.242        | 0.000        | 1.000 1.000 |

Table 3 Results from the reanalysis of the SOC data using the multivariate analysis

Gelman–Rubin $\hat{R}$ values and upper 95% credible bounds near one suggest convergence. $P(\rho > 0)$ is the posterior probability that $\rho$ is greater than zero. Component abbreviations: DD developmental disabilities, MH mental health, JJ juvenile justice, CW child welfare, and ED education system. Using $A = 10$
Due to sparse response, contact with the developmental disabilities and juvenile justice components was not assessed in the original univariate analysis [1]. But using the MVP, we can conduct inference on both. We do not find a significant association in contact with the juvenile justice component ($\rho_{JJ} = -0.021, 95\% \text{ CrI} [-0.081, 0.028], P(\rho_{JJ} > 0) = 0.165$). The relationship between care type and contact with the developmental disabilities component is more nuanced. We find a positive effect suggesting that the marginal probability of contact is $\rho_{DD} = 0.036$ more for primary care than for specialty care. The 95% CrI does include zero, $-0.001, 0.095$, although it is on the edge and the posterior probability that $\rho_{DD}$ is greater than zero is 97.2%. Thus, a sizable portion of the posterior distribution of $\rho_{DD}$ is positive.

Overall, our analysis suggests there are inefficiencies in the system of care framework, a consistent finding with that of Knutson et al. [1]. But our results are more nuanced and suggest multiple areas for targeted improvement. In the study, primary care is recommending contact with both the education and mental health systems at lower rates than they should be; potentially leaving important actors out of the care for children with mental health disorders. Conversely, primary care is contacting disability services more than specialty care; potentially overburdening those services with children who do not need them. Many children only receive mental health care through their primary care physicians [37], so it is important these rates of contact align to ensure the children benefit from the SOC framework [1].

5 Discussion

Methods for MMP data are limited to approaches that do not perform well in the presence of sparse response. In this manuscript, we propose a Bayesian analysis for MMP data that addresses the sparse response in two ways: first by using the scale-invariant latent variable transformation in Eq. (1) which fixes the latent model variance at 1 and second by implementing a hierarchical horseshoe-like prior. The latter imposes a weak bound on the standard deviation of the latent model coefficients that approximately equals the hyper-parameter $A$. To accommodate the multivariate nature of the data, our approach models the latent covariance using a Bayesian FPCA. The Bayesian FPCA gives a low-dimensional expansion using well-behaved basis functions, in this case B-splines, that do not overly burden the analysis with parameters, whereas simulating from a general latent correlation would add a potentially large number of parameters to the model. Prior work demonstrates the ability of the Bayesian FPCA to adequately model a variety of complex covariance structures [28, 29]. The Bayesian FPCA is also useful for modeling latent covariances since those structures may be difficult to observe and therefore hard to pre-simplify with an assumption, e.g., a compound symmetric covariance instead of an unstructured one. Additionally, our approach is more computationally efficient than other multivariate probit models.

As we demonstrate in simulation, the Bayesian MVP outperforms the existing methods in the presence of sparse response regardless of the effect size, as controlled by $\theta_{12}$, the number of outcomes, $K$, or the “true” latent correlation structure.
It routinely has coverage in excess of nominal and does not dip below 91% when constructing 95% intervals. Conversely, the existing methods for MMP data have inconsistent coverage that in some cases can be as low as 84% when the structure is independent and 85% for the block compound symmetric. At the same time, our approach has the highest power without sacrificing absolute bias or coverage and with only minimally larger interval widths. Bias is somewhat higher than expected across all methods. This is to be expected for the Bayesian MVP since the regularization prior makes the bias-variance tradeoff to improve the estimation of the variance. For the remaining methods, it is likely due to the fact that, in the presence of sparse responses, we are often estimating marginal probabilities on the boundary of parameter space.

The reanalysis of the SOC data confirms earlier findings, specifically that contact with the education system is lower among primary care than specialty care. But our approach reveals several additional results that warrant further investigation. Among the components that were previously analyzed, contact with the mental health component may also be lower for primary care than for specialty care. The main advantage our Bayesian MVP provides is in the ability to also assess the components with sparse response which were not examined in the original univariate analysis. What we see is that there is some evidence of increased contact with disability services among primary care compared to specialty care. Taken all together, the SOC study and our reanalysis should provide the foundation for further examination of the coherence of SOC frameworks. These data can be thought of as a pilot study that can provide a good source for developing future, larger studies to further examine the differences in contact within the SOC between primary care and specialty care. Our Bayesian MVP can seamlessly incorporate prior information from this study by replacing the mean of the hierarchical prior on $\beta$ with values derived from the analysis presented in Table 3.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s12561-023-09368-8.

**Acknowledgements** Partial funding for this work was provided by internal Georgetown University Summer Academic Research Grants.

**Declarations**

**Conflict of interest** The authors have no potential or actually conflicts of interest to declare. No additional data were collected for this research. The original study was retrospective in nature and was approved by the Boston University Institutional Review Board [1].

**References**

1. Knutson KH, Meyer MJ, Thakrar N, Stein BD (2018) Care coordination for youth with mental health disorders in primary care. Clin Pediatr 57:5–10. https://doi.org/10.1177/00099228177733740
2. Klingenberg B, Agresti A (2006) Multivariate extension of McNemar’s test. Biometrics 62:921–928. https://doi.org/10.1111/j.1541-0420.2006.00525.x
3. McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12:153–157. https://doi.org/10.1007/BF02295996
4. Consonni G, La Rocca L (2008) Tests based on intrinsic priors for the equality of two correlated proportions. J Am Stat Assoc 103:1260–1269. https://doi.org/10.1198/01621450800000043
5. Saeki H, Tango T, Wang J (2017) Statistical inference for noninferiority of difference in proportions of clustered matched-pair data from multiple raters. J Biopharm Stat 27:70–83. https://doi.org/10.1080/10543406.2016.1148709
6. Westfall PH, Troendle JF, Penaello G (2010) Multiple McNemar tests. Biometrics 66:1185–1191. https://doi.org/10.1111/j.1541-0420.2010.01408.x
7. Xu J, Yu M (2013) Sample size determination and re-estimation for matched pair designs with multiple binary endpoints. Biom J 55:430–443. https://doi.org/10.1002/bimj.201100231
8. Lui K-J, Chang K-C (2013) Testing and estimation of proportion (or risk) ratio under the matched-pair design with multiple binary endpoints. Biom J 55:603–616. https://doi.org/10.1002/bimj.20120224
9. Lui K-J, Chang K-C (2016) Notes on testing noninferiority in multivariate binary data under the matched-pair design. Stat Methods Med Res 25:1272–1289. https://doi.org/10.1177/0962280213477022
10. Cochran WG (1950) The comparison of percentages in matched samples. Biometrika 37:256–266. https://doi.org/10.2307/2332378
11. Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22:719–748. https://doi.org/10.1093/jnci/22.4.719
12. Jiang Y, Xu J (2017) A comparative study of matched pair designs with two binary endpoints. Stat Methods Med Res 26:2526–2542. https://doi.org/10.1177/0962280215601136
13. Agresti A (2013) Categorical data analysis, 3rd edn. Wiley, Hoboken
14. Altham PME (1971) The analysis of matched proportions. Biometrika 58:561–576. https://doi.org/10.2307/2334391
15. Broemeling LD, Gregurich MA (1996) A Bayesian alternative to the analysis of matched categorical responses. Commun Stat 25:1429–1445. https://doi.org/10.1080/03610929608831777
16. Ghosh M, Chen M-H, Ghosh A, Agresti A (2000) Hierarchical Bayesian analysis of binary matched pairs data. Stat Sin 10:647–675
17. Albert J, Chib S (1993) Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc 88:669–679. https://doi.org/10.1080/01621459.1993.10476321
18. Albert J, Chib S (1995) Bayesian residual analysis for binary response models. Biometrika 82:747–759. https://doi.org/10.1093/biomet/82.4.747
19. Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, Cambridge
20. Gelman A (2006) Prior distributions for variance parameters in hierarchical models. Bayesian Anal 1:513–533. https://doi.org/10.1214/06-BA117A
21. Carvalho CM, Polson NG, Scott JG (2010) The horseshoe estimator for sparse signals. Biometrika 97:465–480. https://doi.org/10.1093/biomet/asq017
22. van der Linde A (2008) Variational Bayesian functional PCA. Comput Stat Data Anal 53:517–533. https://doi.org/10.1016/j.csda.2008.09.015
23. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis, 3rd edn. Chapman and Hall-CRC, Boca Raton
24. Chib S, Greenberg E (1998) Analysis of multivariate probit models. Biometrika 85:347–361. https://doi.org/10.1093/biomet/85.2.347
25. Liu C (2001) Discussion. J Comput Graph Stat 10:75–81. https://doi.org/10.1198/10618600152418746
26. Zhang X, Boscardin WJ, Belin TR (2006) Sampling correlation matrices in Bayesian models with correlated latent variables. J Comput Graph Stat 15:880–896. https://doi.org/10.1198/106186006X160050
27. Webb EL, Forster JJ (2008) Bayesian model determination for multivariate ordinal and binary data. Comput Stat Data Anal 52:2632–2649. https://doi.org/10.1016/j.csda.2007.09.008
28. Goldsmith J, Kitago T (2016) Assessing systematic effects of stroke on motor control using hierarchical function-on-scalar regression. J R Stat Soc Ser C 65:215–236. https://doi.org/10.1111/rssc.12115
29. Meyer MJ, Morris JS, Gazes RP, Coull BA (2022) Ordinal probit functional outcome regression with application to computer-use behavior in rhesus monkeys. Ann Appl Stat 16:537–550. https://doi.org/10.1214/21-AOAS1513
30. Gupta AK, Nagar DK (2000) Matrix variate distributions, 2nd edn. Chapman & Hall/CRC, Boca Raton
31. Eilers PHC, Marx BD (1996) Flexible smoothing with b-splines and penalties. Stat Sci 11:89–121. https://doi.org/10.1214/ss/1038425655
32. Polson NG, Scott JG (2012) On the half-Cauchy prior for a global scale parameter. Bayesian Anal 7:887–902. https://doi.org/10.1214/12-BA730
33. Wand MP, Ormerod JT, Padoan SA, Frühwirth R (2011) Mean field variational bayes for elaborate distributions. Bayesian Anal 6:847–900. https://doi.org/10.1214/11-BA631
34. Agresti A, Coull BA (1998) Approximate is better than ‘exact’ for interval estimation of binomial proportions. Am Stat 52:119–126. https://doi.org/10.1080/00031305.1998.10480550
35. Agresti A, Caffo B (2000) Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. Am Stat 54:280–288. https://doi.org/10.1080/00031305.2000.10474560
36. Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7:457–511. https://doi.org/10.1214/ss/1177011136
37. Costello EJ, He J-P, Sampson NA, Kessler RC, Merikangas KR (2014) Services for adolescents with psychiatric disorders: 12-month data from the National Comorbidity Survey-Adolescent. Psychiatr Serv 65:359–366. https://doi.org/10.1176/appi.ps.201100518