Gram–Schmidt–Fisher scoring algorithm for parameter orthogonalization in MLE

John Kwagyan1*, Victor Apprey1 and †George E. Bonney1

Abstract: The estimation of parameters is a key component in statistical modelling and inference. However, parametrization of certain likelihood functions could lead to highly correlated estimates, causing numerical problems, mathematical complexities and difficulty in estimation or erroneous interpretation and subsequently inference. In statistical estimation, the concept of orthogonalization is familiar as a simplifying technique that allows parameters to be estimated independently and thus separates information from each other. We introduce a Fisher scoring iterative process that incorporates the Gram–Schmidt orthogonalization technique for maximum likelihood estimation. A finite mixture model for correlated binary data is used to illustrate the implementation of the method with discussion of application to oesophageal cancer data.

1. Introduction

The problem of estimating parameters is one of the key stages in fitting a statistical model to a set of data. Typically, the model will contain some parameters which are not of interest in themselves, but whose values affect the inferences that we make about the parameters of direct interest. Parametrization used when working with certain distributions could lead to high correlations of the maximum likelihood estimates. Maximization algorithms converge rapidly if the initial estimates are good, the likelihood function is well approximated by a quadratic in the neighbourhood of the parameter space and the information matrix is well conditioned, which means that the parameter estimates are not strongly intercorrelated. High intercorrelation of parameters causes numerical problems, and difficulty or erroneous interpretation in the parameter estimation. In statistical

ABOUT THE AUTHOR

Dr John Kwagyan is a mathematician and a public health and medical researcher. His research interests include mathematical and statistical modelling of correlated data, clinical trials, survival models, big data analytics, statistical genetics and pharmacokinetic & pharmacodynamics modeling. He is currently the director of Biostatistics, Epidemiology and Research Design, of the Georgetown-Howard University Center for Clinical and Translational Science, Howard University College of Medicine and adjunct professor in the Department of Mathematics, Howard University, Washington, DC, USA.

PUBLIC INTEREST STATEMENT

In statistical estimation, the concept of orthogonalization is simply a technique that allows parameters to be estimated independently, through the rotation of the parameter space. The paper introduces an iterative algorithm that utilizes the Fisher scoring process and incorporates the Gram–Schmidt orthogonalization technique for maximum likelihood estimation. The algorithm that we propose provides an exceedingly convenient method for multi-parameter statistical problems first to reduce numerical difficulties and second to improve accuracy in parameter estimation.
estimation, the concept of orthogonalization is familiar as a simplifying technique that allows parameters to be estimated independently, through rotation of the parameter space. Computationally, all that is required is that the original parameters should be computable from the new ones, and vice-versa. An injective transformation will therefore be desirable, though not necessary.

Ross (1970) gave a comprehensive discussion of techniques that can be used to reduce correlation in particular problems. Among them are sequential and nested maximizations. Karlbliesch and Sprott (1970) discuss methods aimed at eliminating nuisance parameters from the likelihood function so that inference can be made about only the parameters of interest. Amari (1982, 1985) derived from a theoretical point of view, a general criteria for the existence and construction of orthogonal parameters for statistical models using differential geometry. Cox and Reid (1987) gave a more general procedure and covered the advantages in maximum likelihood estimation. The approach, though popular in theoretical statistics, is computationally impractical in many application situations. Willmot (1988) discussed orthogonal parametrization but only for a two-parameter family of discrete distributions. His method, however, does not allow for higher parametric problems. Hurlimann (1992) discusses the existence of orthogonal parameterization to the mean, characterized by a partial differential equation involving the mean, the variance and cumulant generating functions. Godambe (1991) deals with the problem of nuisance parameters, within a semi-parametric set-up which includes the class of distributions associated with generalized linear models. Their technique uses the optimum orthogonal estimating functions of Godambe and Thompson (1989). Bonney and Kissling (1986) applied the Gram–Schmidt orthogonalization (Clayton, 1971) to multivariate data and presented an application in genetics. Kwagyan, Apprey and Bonney (2001) presented the idea of Gram–Schmidt parameter orthogonalization in genetic analysis. A major consequence of parameter orthogonalization is that the maximum likelihood estimate of one parameter is not affected or changes slowly with the misspecification of the other. The most direct interpreted consequence is that the maximum likelihood estimates of orthogonal coordinates are asymptotically independent.

The remainder of this paper is organized as follows. In Section 2, the disposition model for clustered binary data proposed by Bonney (1995, 1998a) and further investigated by Kwagyan (2001) is introduced and used as a motivation. In Section 3, an overview of maximum likelihood estimation and parameter orthogonalization is then presented. Section 4 develops the proposed Gram–Schmidt parameter orthogonalization and iterative scheme for maximum likelihood estimation. In section Section 5, fitting algorithm of the disposition model is discussed with application to oesophageal cancer data in Chinese families. Finally, Section 6 wraps up with discussion of the methodology and outlines directions for further research.

2. Model for correlated binary data
Consider a sample of \( N \) groups, each of size \( n_i, i = 1, \ldots, N \); and let \( Y_i = (Y_{i1}, \ldots, Y_{in_i})^T \) denote the vector of binary outcomes for the \( i \)th group. It is postulated that the measures of the outcomes within a group are possibly correlated. Further suppose the \( j \)th subject in the \( i \)th group has \( 1 \times p \) individual-specific covariates \( X_{ij} = (x_{ij1}, \ldots, x_{ijp}) \) and let the \( i \)th group have group-specific covariates \( Z_i = (z_1, \ldots, z_q) \). Let \( \delta_{ij} \) denote the conditional probability of \( Y_{ij} = 1 \) given \( Y_{ij'} = 1 \); that is

\[
\delta_{ij} = P(Y_{ij} = 1|Y_{ij'} = 1), \quad j \neq j', \quad j, j' = 1, 2, \ldots, n.
\]

We call \( \delta_{ij} \) the individual disposition which is simply interpreted as the probability of the outcome on one unit given another unit from the same cluster has the attribute. This in essence is an indication of aggregation. Assume further that within the group, a pair of observations satisfy the relation

\[
\frac{P(Y_{ij} = 1, Y_{ij'} = 1)}{P(Y_{ij} = 1) P(Y_{ij'} = 1)} = \frac{1}{\alpha_i}, \quad \alpha_i < 0, \quad j \neq j', \quad j, j' = 1, 2, \ldots, n,
\]
where $\alpha_i$ is common for all pairs. Clearly, $\alpha_i = 1$ implies independence of the observations. Thus, $\alpha_i$ is a measure of the departure from independence and is called the relative disposition. With the above definitions, Bonney (1995, 1998a), and from further investigation through a latent mixture formulation by Kwagyan (2001), has shown that the joint distribution for the $N$ groups can be based on

$$L(\Theta) = \prod_{i=1}^{N} \left\{ (1 - \alpha_i) \prod_{j=1}^{n_i} (1 - y_{ij}) + \alpha_i \prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}} (1 - \delta_{ij})^{1-y_{ij}} \right\}$$  \hspace{1cm} (2.1)

The $\alpha_i$ and $\delta_{ij}$ are generally modelled as

$$\alpha_i = \frac{1 + e^{-\langle M(Z_i), D(Z_i) \rangle}}{1 + e^{-\langle M(Z_i), D(Z_i) \rangle + W(X_i)}} \quad \text{and} \quad \delta_{ij} = \frac{1}{1 + e^{-\langle M(Z_i), D(Z_i) \rangle + W(X_i)}}$$

where $M(Z_i)$ is function of mean effects, $D(Z_i)$ is a measure of within-group dependence and $W(X_i)$ a function describing the effect of the individual covariates. Typically, $M(Z_i)$, $D(Z_i)$ and $W(X_i)$ are modelled, respectively, as

$$M(Z_i) = \lambda_0 + \lambda_1 Z_1 + \ldots + \lambda_p Z_p$$
$$D(Z_i) = \gamma_0 + \gamma_1 Z_1 + \ldots + \gamma_q Z_q$$
$$W(X_i) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_r X_r$$

where

$$\Theta' = \{ \Lambda, \Gamma, \beta \}$$

and

$$\Lambda = \{ \gamma_0, \ldots, \gamma_q \} \quad \Gamma = \{ \lambda_0, \ldots, \lambda_p \} \quad \beta = \{ \beta_1, \ldots, \beta_r \}$$

are unknown parameters to be estimated. When data consist of a few large clusters or truncated or size-biased samples, Bonney (1998b) has shown that there is little or no information to separate the effects of $M(Z)$ and $D(Z)$. The estimates of the parameters tend to be highly correlated. Similar problems occur in other approaches as well. The popular estimating equations approach, GEE (Liang & Zeger, 1986), suffers the same problems when there are only a few large clusters (see discussion of Prentice, 1988).

### 3. Maximum likelihood estimation and parameter orthogonalization

There are different statistical methods for estimating parameters, but the approach most commonly used is that based on maximum likelihood estimation. Maximum likelihood estimates of parameters are those values of the parameters that make the likelihood function a maximum. Let $y = (y_1, y_2, \ldots, y_n)$ be observed data from a population with a probability distribution $P(y; \theta)$ indexed by the vector of unknown parameters $\theta = (\theta_1, \ldots, \theta_p)$. The likelihood function for $\theta$ is the joint distribution

$$L(\theta; y) = \prod_{i=1}^{n} P(y_i; \theta)$$

According to the likelihood principle, $\hat{\theta}$ is regarded as the value of $\theta$ which is most strongly supported by the observed data. In practice, $\hat{\theta}$ is obtained by direct maximization of $\log L(\theta)$ or by solving the set of equations from the score function, $U(\theta)$, where

$$U(\theta) = \frac{\partial \log L}{\partial \theta} = 0$$
for which the Hessian matrix $H(\theta) = \left( \frac{\partial^2 \log L}{\partial \theta \partial \theta} \right)$ is negative definite. The matrix $I(\theta) = E(-H(\theta)) = E(-\frac{\partial^2 \log L}{\partial \theta \partial \theta})$ is called Fisher (expected) Information matrix for $\theta$ and its inverse, $\Omega(\theta) = [I(\theta)]^{-1}$ gives the asymptotic variances and covariances of the maximum likelihood estimates.

Suppose $\theta$ is a $p$ parameter vector that is partitioned into two vectors $\theta_1$ and $\theta_2$ of lengths $p_1$ and $p_2$, respectively, $p_1 + p_2 = p$. Cox and Reid (1987) define $\theta_1$ to be orthogonal to $\theta_2$, if the elements of the information matrix satisfy

$$E \left( \frac{\partial^2 l}{\partial \theta_s \partial \theta_t} \right) = \left( -\frac{\partial^2 l}{\partial \theta_s \partial \theta_t} \right) = 0, \quad s = 1, \ldots, p_1; \quad t = p_1 + 1, \ldots, p$$

If this holds for all $\theta$ in the parameter space, it is sometimes called global orthogonality. If it holds at only one parameter value say $\theta^0$, then $\theta_1$ and $\theta_2$ are said to be locally orthogonal at $\theta^0$.

We now discuss Cox and Reid (1987) approach for the construction of orthogonal parameters. Suppose that initially the likelihood is specified in terms of $(\theta, \gamma)$, $\gamma = (\gamma_1, \ldots, \gamma_p)$ and let $l(\theta, \gamma)$ be the log-likelihood function. Assume $\theta$ is the parameter of interest and $\gamma$ the set of nuisance parameters. We seek a transformation from $(\theta, \gamma)$ to a new set of orthogonal parameters $(\theta, \lambda)$, $\lambda = (\lambda_1, \ldots, \lambda_p)$. It is easiest to think of the original parameters $\gamma$ as some function of the new parameters, $\lambda$, that is

$$\gamma = \gamma(\theta, \lambda)$$

Then, the log-likelihood function can be expressed as

$$l'(\theta, \lambda) = l(\theta, \gamma_1(\theta, \lambda), \ldots, \gamma_p(\theta, \lambda)),$$

where now $l'$ refers to the log-likelihood in the new parametrization. Taking derivatives of this equation with respect to the new parameters, we have by the chain rule;

$$\frac{\partial l'}{\partial \theta} = \frac{\partial l}{\partial \theta} + \sum_{j=1}^{p} \frac{\partial l}{\partial \gamma_j} \frac{\partial \gamma_j}{\partial \theta}$$

and

$$\frac{\partial^2 l'}{\partial \theta \partial \lambda_k} = \sum_{i=1}^{p} \frac{\partial^2 l}{\partial \theta \partial \gamma_i} \frac{\partial \gamma_i}{\partial \lambda_k} + \sum_{j=1}^{p} \frac{\partial l}{\partial \gamma_j} \frac{\partial^2 \gamma_j}{\partial \theta \partial \lambda_k} + \sum_{i=1}^{p} \frac{\partial l}{\partial \gamma_i} \frac{\partial^2 \gamma_i}{\partial \theta \partial \lambda_k} \quad (3.1)$$

The derivatives of $\gamma$ are not functions of the data, and hence are constant with respect to expectation. Therefore, after taken expectations with respect to the distribution of the data indexed by the parameters $\{\theta, \gamma\}$, the last term in Equation 3.1 is zero. If the parameters are orthogonal, then again from Equation 3.1, we have

$$\sum_{j=1}^{p} \frac{\partial \gamma_j}{\partial \lambda_k} \left( E \left( \frac{\partial^2 l}{\partial \theta \partial \gamma_j} \right) + \sum_{j=1}^{p} E \left( \frac{\partial^2 l}{\partial \gamma_k \partial \gamma_j} \right) \frac{\partial \gamma_j}{\partial \theta} \right) = 0.$$ 

so that the orthogonality equations are

$$\sum_{j=1}^{p} \frac{\partial \gamma_j}{\partial \lambda_k} \left( E \left( \frac{\partial^2 l}{\partial \theta \partial \gamma_j} \right) + \sum_{j=1}^{p} E \left( \frac{\partial^2 l}{\partial \gamma_k \partial \gamma_j} \right) \frac{\partial \gamma_j}{\partial \theta} \right) = 0, \quad k = 1, \ldots, p.$$ 

We require the transformation from $(\theta, \gamma)$ to $(\theta, \lambda)$ to have a non-zero Jacobian; hence,

$$\sum_{j=1}^{p} E \left( \frac{\partial^2 l}{\partial \gamma_j \partial \gamma_k} \right) \frac{\partial \gamma_j}{\partial \theta} = -E \left( \frac{\partial^2 l}{\partial \theta \partial \gamma_k} \right), \quad k = 1, \ldots, p.$$ 

Page 4 of 15
This is a system of $p$ differential equations which must be solved for $\gamma(\theta, \lambda)$. In fact, since $\lambda$ does not enter explicitly into the equations, the solution for $\gamma$ can contain an arbitrary function of $\lambda$ as the integrating constant. It is noted in the discussion of Cox and Reid that although it is sometimes theoretically possible to find a differential equation, simple explicit solutions of the differential equation were not feasible for the same models. There are also cases where explicit solutions are possible, but the original nuisance parameters could not be explicitly expressed in terms of the orthogonal parameter (Hills, 1987). In general, global orthogonalization can also not be achieved by this approach.

4. Gram–Schmidt parameter orthogonalization

The Gram–Schmidt orthogonalization process (Clayton, 1971) is equivalent to the linear transformation $\Theta = (\theta_1, \ldots, \theta_p)$ to $\Theta^* = (\theta^*_1, \ldots, \theta^*_p)$ defined as:

$$
\theta^*_1 = \theta_1 \\
\theta^*_2 = \theta_2 - b_{21} \theta_1 \\
\vdots \\
\theta^*_j = \theta_j - \sum_{k=1}^{j-1} b_{j,k} \theta^*_k, j = 2, \ldots, p
$$

(4.1)

In linear regression set-up, $\theta^*_j$ is $\theta_j$ adjusted for $\theta_1, \ldots, \theta_{j-1}$, and $b_{p,k}$, $j = 2, \ldots, p; k = 1, \ldots, (j - 1)$ are the multiple regression coefficients. Writing this using matrix notation, we have

$$
\Theta^* = B \Theta
$$

where $B$, the transformation matrix, is lower triangular with ones along the diagonal. The transformation matrix is chosen so that $\theta^*_1, \theta^*_2, \ldots, \theta^*_p$ are mutually uncorrelated. The Jacobian of the transformation is unity. Suppose $\Omega$ is the covariance matrix of $\Theta$, then the covariance matrix of $\Theta^*$, $\Omega^* = B \Omega B^T$ is a diagonal matrix. We recommend that the parameters be ordered in terms of interest. This ensures that the parameters of most interest are least affected by round-off errors.

4.1. Evaluation of the transformation matrix

To evaluate the transformation matrix, $B$, let $\Sigma = (c_{ij})$, $i, j = 1, \ldots, p$, where

$$
c_{ij} = \text{cov}(\theta_i, \theta_j) = \left[ E \left( -\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right) \right],
$$

are elements of the covariance matrix. The orthogonality relation, $c^*_i = \text{cov}(\theta^*_i, \theta^*_j) = 0$ ($i \neq j$) implies that for $i < j$,

$$
0 = c^*_j = \text{cov}(\theta^*_i, \theta^*_j) = \text{cov}(\theta_i, \theta_j) \sum_{k=1}^{j-1} b_{j,k} \text{cov}(\theta_i, \theta_{j-k})
$$

$$
= \text{cov}(\theta_i, \theta_j) \sum_{k=1}^{j-1} b_{j,k} \left\{ \text{cov}(\theta_{j-k}, \theta_j) + \sum_{k' = 1}^{j-1} b_{j-k', k} \text{cov}(\theta_{j-k'}, \theta_{j-k}) \right\}
$$

$$
= \text{cov}(\theta_i, \theta_j) \sum_{k=1}^{j-1} b_{j,k} \left\{ c_{j-k,j} + \sum_{k' = 1}^{j-1} b_{j-k', k} c_{j-k', j-k} \right\}
$$

$$
= \text{cov}(\theta_i, \theta_j) \sum_{k=1}^{j-1} b_{j,k} c_{j-k,j} - \sum_{k=1}^{j-1} b_{j,k} \left\{ c^*_j - \sum_{k' = 1}^{j-1} b_{j-k', k} c^*_{j-k', j-k} \right\}
$$

$$
= \text{cov}(\theta_i, \theta_j) \sum_{k=1}^{j-1} b_{j,k} c_{j-k,j} - \sum_{k=1}^{j-1} b_{j,k} c^*_{j-k,j}
$$

$$
= \text{cov}(\theta_i, \theta_j) \sum_{k=1}^{j-1} b_{j,k} c_{j-k,j} - \sum_{k=1}^{j-1} b_{j,k} c^*_{j-k,j}, \text{ since } c^*_{j-k,j} = 0
$$
Thus, the system of linear equations to be solved for entries of the transformation matrix based on the covariance matrix is:

\[ c_j = \sum_{k=1}^{j-1} b_{j-k} c_{j-k}, \quad i = 1, \ldots, (j - 1), \quad j = 2, \ldots, p. \] (4.2)

Consequently, solving the system of linear equations (4.2), we obtain the elements of the transformation matrix as

\[ b_{jk} = \sum_{r=1}^{j-1} c_r I_{r(j-k)}, \quad j = 2, \ldots, p; \quad k = 1, \ldots, (j - 1). \] (4.3)

where \( I_{r(j-k)} \) are entries of the information matrix. The observed information matrix would be a good approximation of the expected information, if there is difficulty evaluating it.

For illustration, when \( p = 4 \), we have Gram–Schmidt orthogonalization process in matrix notation as,

\[
\begin{pmatrix}
\theta_1^* \\
\theta_2^* \\
\theta_3^* \\
\theta_4^*
\end{pmatrix} =
\begin{pmatrix}
1 & 0 & 0 & 0 \\
-b_{21} & 1 & 0 & 0 \\
-b_{31} & -b_{32} & 1 & 0 \\
-b_{41} & -b_{42} & -b_{43} & 1
\end{pmatrix}
\begin{pmatrix}
\theta_1 \\
\theta_2 \\
\theta_3 \\
\theta_4
\end{pmatrix}
\]

The orthogonalization relationship of the parameters \((\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*)\) implies

\[
\begin{align*}
0 &= c_{12} - b_{21} c_{11} \\
0 &= c_{13} - b_{32} c_{12} - b_{31} c_{11} \\
0 &= c_{23} - b_{32} c_{22} - b_{31} c_{12} \\
0 &= c_{14} - b_{43} c_{13} - b_{42} c_{12} - b_{41} c_{11} \\
0 &= c_{24} - b_{43} c_{23} - b_{42} c_{22} - b_{41} c_{12} \\
0 &= c_{34} - b_{43} c_{33} - b_{42} c_{23} - b_{41} c_{13}
\end{align*}
\]

And so the system of linear equations to be solved for the elements of the transformation matrix is:

\[
\begin{pmatrix}
c_{11} & 0 & 0 & 0 \\
0 & c_{11} & c_{12} & 0 \\
0 & c_{12} & c_{22} & 0 \\
0 & 0 & 0 & c_{11} \\
0 & 0 & 0 & c_{12} \\
0 & 0 & 0 & c_{22} \\
0 & 0 & 0 & c_{13} \\
0 & 0 & 0 & c_{23} \\
0 & 0 & 0 & c_{33}
\end{pmatrix}
\begin{pmatrix}
b_{21} \\
b_{31} \\
b_{41} \\
b_{22} \\
b_{32} \\
b_{42} \\
b_{23} \\
b_{33} \\
b_{43}
\end{pmatrix}
= \begin{pmatrix}
c_{12} \\
c_{13} \\
c_{23} \\
c_{14} \\
c_{24} \\
c_{34}
\end{pmatrix}
\]

(4.4)

Let \( Q \) be the matrix of coefficients for the system of equations (4.4) Above; then, we note that \( Q \) is a patterned block diagonal matrix. Furthermore, let \( D_r (r = 1, 2, 3) \), be the block diagonal of \( Q \); then, in this illustration,

\[
Q = \begin{pmatrix}
D_1 & 0 & 0 \\
0 & D_2 & 0 \\
0 & 0 & D_3
\end{pmatrix}
\]
where

\[ D_1 = (c_{11}), D_2 = \begin{pmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{pmatrix}, \quad D_3 = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{12} & c_{22} & c_{23} \\ c_{13} & c_{23} & c_{33} \end{pmatrix} \]

Thus, \( D_r \) is the covariance matrix of the first \( r \) parameters. It can easily be shown that \( D_r \) is symmetric and positive definite and so \( Q \) is non-singular. A unique solution thus exists for the system of linear Equation (4.3) and in general for Equation (4.2). And so, the elements of the transformation matrix, \( B \), can be obtained as

\[
\begin{pmatrix}
 b_{11} \\
 b_{31} \\
 b_{21} \\
 \vdots \\
 b_{43}
\end{pmatrix} = \begin{pmatrix}
 c_{11} & 0 & 0 & 0 & 0 \\
 0 & c_{11} & c_{12} & 0 & 0 \\
 0 & c_{12} & c_{22} & 0 & 0 \\
 0 & 0 & 0 & c_{11} & c_{12} \\
 0 & 0 & 0 & c_{12} & c_{22} \\
 0 & 0 & 0 & c_{13} & c_{23}
\end{pmatrix}^{-1} \begin{pmatrix}
 c_{12} \\
 c_{13} \\
 c_{23} \\
 c_{14} \\
 c_{24} \\
 c_{34}
\end{pmatrix}
\]

or

\[
\begin{pmatrix}
 b_{11} \\
 b_{31} \\
 b_{21} \\
 \vdots \\
 b_{43}
\end{pmatrix} = \begin{pmatrix}
 I_{11} & 0 & 0 & 0 & 0 \\
 0 & I_{11} & I_{12} & 0 & 0 \\
 0 & I_{12} & I_{22} & 0 & 0 \\
 0 & 0 & 0 & I_{11} & I_{12} \\
 0 & 0 & 0 & I_{12} & I_{22} \\
 0 & 0 & 0 & I_{13} & I_{23}
\end{pmatrix} \begin{pmatrix}
 c_{12} \\
 c_{13} \\
 c_{23} \\
 c_{14} \\
 c_{24} \\
 c_{34}
\end{pmatrix}
\]

where \( I_j \) are the entries of the information matrix.

Having obtained \( \Theta^* = (\theta_1^*, \ldots, \theta_p^*) \) the original parameters, \( \Theta = (\theta_1, \ldots, \theta_p) \) can be obtained recursively as

\[ \theta_j = \theta_j^* + \sum_{k=1}^{j-1} b_{j-k} \theta_{j-k}, \quad j = 2, \ldots, p \]

or writing this using matrix notation, we have

\[ \Theta = B^{-1} \Theta^* \]

4.2. Block orthogonalization

Let \( \Theta^T = (\Lambda, \Gamma, \beta) \) be a set of parameters to be estimated where

\[ \Lambda = (\lambda_0, \lambda_1, \ldots, \lambda_q)^T; \quad \Gamma = (\gamma_0, \gamma_1, \ldots, \gamma_p)^T; \quad \beta = (\beta_1, \beta_1, \ldots, \beta_p)^T \]

Suppose further that the set \( \Theta^* = (\Lambda^*, \Gamma^*, \beta^*) \) is correlated. Then, we wish to find a new set \( \Theta^{*T} = (\Lambda^T, \Gamma^T, \beta^T) \) through a linear transformation such that the vector of parameters in \( \Theta^{*T} \) is mutually uncorrelated. We allow for correlation, if any, within each set of parameters in \( \Omega^* \).
Further suppose the vector $\beta$ is the set of parameters of interest. Then, the Gram–Schmidt orthogonalization process computes the new set of parameters in terms of the original parameters as follows:

$$\beta^* = \beta$$
$$\Lambda^* = \Lambda - B_{21} \beta$$
$$\Gamma^* = \Gamma - B_{31} \Lambda - B_{32} \beta$$

where $B_{21}, B_{31}$ and $B_{32}$ are matrices with dimensions $q \times r$, $p \times q$ and $p \times r$, respectively.

In matrix notation, we write

$$\begin{pmatrix}
\beta^* \\
\Lambda^* \\
\Gamma^*
\end{pmatrix} =
\begin{pmatrix}
1 & 0 & 0 \\
-B_{21} & 1 & 0 \\
-B_{31} & -B_{32} & 1
\end{pmatrix}
\begin{pmatrix}
\beta \\
\Lambda \\
\Gamma
\end{pmatrix}$$

or

$$\Theta^* = B \Theta$$

where $B$ is a lower triangular block matrix whose diagonal unit matrix is the transformation matrix chosen such that $\Lambda^*, \Gamma^*, \beta^*$ are mutually uncorrelated and where $\Theta^T \Theta$ is a block diagonal matrix. The Jacobian of the transformation is unity. The only way in which this procedure could break down would be if one of the vectors of $\Theta$ is identically zero. From the method of formation of $\Theta^*$, it is clear that it is a linear combination of the vectors in $\Theta$. If $\Theta$ is identically zero, this means that $\Theta$ are linearly dependent, thus contradicting the assumption of independence. Extensions to more than three sets of vectors of parameters readily follow.

### 4.3. Gram–Schmidt–Fisher scoring algorithm

We introduce a modification of the Fisher scoring algorithm incorporating the Gram–Schmidt process to obtain an information matrix which is diagonal and thus ensuring the approximate (near) orthogonality and consequently the stability of the estimates of the new parameters. Since the transformation is linear and injective, the original parameters are readily obtained.

Let $l(\Theta; y)$ be the log-likelihood in the original Parametrization; then, Fisher scoring algorithm is given by the iterative routine

$$\theta_{m+1} = \theta_m + [I(\theta_m)]^{-1}U(\theta_m)$$

(4.5)

where $I(\theta)$ is the expected information matrix and $U(\theta)$ is the score function.

Suppose $\Theta$ is transformed to $\Theta^*$ through the Gram–Schmidt orthogonalization process. Let $B = (b_{ij})$ be the matrix of transformation, defined as in Equation (4.3). Then, since $B$ is square and non-singular, we can write from Equation (4.5).

$$B\theta_{m+1} = B\theta_m + B[I(\theta_m)]^{-1}B^T(B^T)^{-1}U(\theta_m)$$

$$B\theta_{m+1} = B\theta_m + [B[I(\theta_m)]^{-1}B^T][B^{-1}U(\theta_m)]$$

$$\theta_{m+1}^* = \theta_{m+1}^* + [I(\theta_m^*)]^{-1}U(\theta_{m+1}^*)$$

(4.6)

where

$$\theta_m^* = B\theta_m$$

$I(\theta_m^*) = B[I(\theta_m)^{-1}]B^T$ is asymptotically diagonal

$U(\theta_m^*) = (B^T)^{-1}U(\theta_m)$

We shall call Equation (4.6) the Gram–Schmidt–Fisher scoring algorithm.
4.3.1. Algorithmic process

The proposed Gram–Schmidt–Fisher scoring algorithm is a 2-stage iterative process that oscillates between Equations (4.3) and (4.6) until convergence and is described iteratively as follows:

1. Start with an initial estimate of the original parameter, \( \theta_0 \), denoted by \( \theta^{(0)} \).
2. Estimate \( B \) at \( \theta^{(0)} \) from Equation 4.3 that is \( b_{jk} = \sum_{i=1}^{n-1} c_{ij} \beta_{ij-k} \), \( j = 2, \ldots, p; \ k = 1, \ldots, (j-1) \).
3. Determine the initial estimate of the orthogonal parameterization, \( \theta^{(1)} \) from Equation 4.1, i.e.
   \[ \theta^{(1)} = \theta^{(0)} - \sum_{k=1}^{j-1} b_{jk} \theta^{(0)} \]
4. Update \( \theta^* \) to obtain a new value for \( \theta^{(1)} \), that is \( \theta^{(1)} = \theta^{(0)} + \theta^* \).
5. Update \( \theta \) to obtain a new value for \( \theta^{(1)} \), recursively as \( \theta^{(1)} = B^{-1} \theta^{(1)} \).
6. Repeat steps 2 through 5 using \( \theta^{(n-1)} \) to obtain \( \theta^{(n)} \).
7. Stop when \( |\theta^{(n-1)} - \theta^{(n)}| < \epsilon \).

4.4. Example

We consider a sample problem which concerns inference about the difference between two exponential means. Let \( Y_1 \) and \( Y_2 \) be the independent exponential random variables with means \( \phi \) and \( \phi + \psi \), respectively. Then, the joint distribution is given by the function

\[ f(Y_1, Y_2 | \phi, \psi) = \frac{1}{\phi (\phi + \psi)} \exp \left( -\frac{Y_1}{\phi} - \frac{Y_2}{\phi + \psi} \right) \]

The score vector is

\[ \mathbf{U}(\phi, \psi) = \left( \frac{\partial}{\partial \phi} \frac{\partial}{\partial \psi} \right) \left( \begin{array}{c} \frac{n}{\phi} - \frac{n}{\phi + \psi} + \frac{\sum_{i=1}^{n} Y_i}{\phi^2} + \frac{\sum_{i=1}^{n} Y_i}{(\phi + \psi)^2} \\ -\frac{n}{\phi} + \frac{n}{\phi + \psi} + \frac{\sum_{i=1}^{n} Y_i}{\phi^2} + \frac{\sum_{i=1}^{n} Y_i}{(\phi + \psi)^2} \end{array} \right) \]

The information matrix is

\[ I_{\phi \psi} = \left( \begin{array}{cc} i_{\phi \phi} & i_{\phi \psi} \\ i_{\psi \phi} & i_{\psi \psi} \end{array} \right) = \left( \begin{array}{cc} \frac{n}{\phi (\phi + \psi)^2} & \frac{n}{(\phi + \psi)^2} \\ \frac{n}{(\phi + \psi)^2} & \frac{n}{\phi (\phi + \psi)^2} \end{array} \right) \]

and its inverse, the variance–covariance matrix, is

\[ I^{-1}_{\phi \psi} = \left( \begin{array}{cc} i_{\phi \phi} & i_{\phi \psi} \\ i_{\psi \phi} & i_{\psi \psi} \end{array} \right) = \left( \begin{array}{cc} \frac{1}{n} \left( 2 \phi^2 + 2 \phi \psi + \psi^2 \right) & -\frac{1}{n} \phi^2 \\ -\frac{1}{n} \phi^2 & \frac{1}{n} \psi^2 \end{array} \right) \]

4.4.1. Cox and Reid approach

Orthogonal parametrization, following Cox–Reid method, requires solving the differential equation

\[ \left( \frac{1}{(\phi + \psi)} + \frac{1}{\phi^2} \right) \frac{\partial \phi}{\partial \psi} = \frac{1}{\phi + \psi} \]

This can be solved by separation of variables, leading to \( a(\lambda) = \phi |(\psi + 2\phi) \), where \( a(\lambda) \) is an arbitrary function of \( \lambda \). Cox–Reid suggest setting \( a(\lambda) = e^\lambda \) as a suitable choice. Clearly, this does not produce a unique solution for \( \phi \), regardless of the parametrization of \( a(\lambda) \). Thus, different reparametrizations may lead to different modified likelihoods, so that a Cox–Reid estimator may not exist or there may be many of them.

4.4.2. Gram–Schmidt approach

The proposed Gram–Schmidt approach requires seeking a transformation \((\psi, \phi) \) to \((\lambda, \phi(\psi, \phi)) \) such that \( \lambda = \phi - b_{21} \psi \) and solving the equation
From the variance–covariance matrix, we obtain

\[ \nu = b_{21} \nu \]

Just like the Cox–Ried approach, it is impractical to obtain or express \( \phi \) uniquely in terms of \( \lambda \) and \( \psi \) and subsequently formulate a modified approximate orthogonal joint distribution function. Therefore, one could proceed iteratively, using the proposed Gram–Schmidt–Fisher scoring algorithm.

5. Fitting the disposition model

We will present procedures for estimating the parameters in the correlated model Equation (2.1). The calculations are standard, and so we will only outline the results. We write the likelihood of the joint distribution of the \( i \)-th group as

\[ L_i(Y_i; \theta) = (1 - \alpha_i) \prod_{j=1}^{n_i} (1 - y_{ij}) + \alpha_i L_{0i}(\theta|y_i) \]

where

\[ L_{0i}(\theta|y_i) = \prod_{j=1}^{n_i} \delta_{ij}^{y_i}(1 - \delta_{ij})^{1 - y_i} \]

\[ \delta_{ij}(\theta) = \frac{1}{1 + \exp[-(M(Z_i) + D(Z_{ij}) + W(X_{ij}))]} \]

\[ a_i(\theta) = \frac{1 + \exp[-(M(Z_i) + D(Z_{ij})]}}{1 + \exp[-M(Z_i)]} \]

Let

\[ Y_{ij}(\theta) = M(Z_i) + D(Z_{ij}) + W(X_{ij}) \]

and define the following

\[ Y^{(1)}_{ij} = \frac{\partial}{\partial \theta} Y_{ij}(\theta) \text{ and } Y^{(2)}_{ij} = \frac{\partial^2}{\partial \theta^2} Y_{ij}(\theta) \]

The contribution of the \( i \)-th group to the log-likelihood is the term

\[ \log L_i(\theta|y_i) = \log \left( (1 - \alpha(\theta)) \prod_{j=1}^{n_i} (1 - y_{ij}) + \alpha(\theta) L_{0i}(\theta|y_i) \right) \]

and the contribution to the score function is the term

\[ U_i = \frac{\partial}{\partial \theta} \log L_i(\theta) = A_i(\theta) \alpha_i + D_i(\theta) U_{0i}(\theta|y), \]

where
The asymptotic variance–covariance matrix of the parameter estimates, \( \mathbf{C}(\theta) = [\mathbf{I}(\theta)]^{-1} \), is the inverse of the information matrix. Thus, the transformation matrix for use of the proposed Gram–Schmidt–Fisher scoring algorithm can be obtained as described from Equation 4.4. Specifically, the transformation matrix, a lower triangular matrix with ones along the diagonal, is \( \mathbf{B} = (\mathbf{b}_{jk}) \), where \( \mathbf{b}_{jk} = \sum_{i=1}^{n} c_{ij} \mathbf{I}_{i}(\mathbf{b}) \). The estimates can be alternatively obtained by the Fisher scoring method which is given by

\[
\theta_{s+1} = \theta_{s} + [\mathbf{I}(\theta_{s})]^{-1} \mathbf{U}(\theta_{s})
\]

and where \( \mathbf{A}_{s}, \mathbf{D}_{s} \), and \( \mathbf{U}_{s} \) are evaluated at \( \mathbf{y} = (y_{1}, y_{2}, \ldots, y_{n}) = \mathbf{0} \). The estimates can be alternatively obtained by the Fisher scoring method which is given by

\[
\theta_{s+1}^{*} = \theta_{s}^{*} + [\mathbf{I}(\theta_{s}^{*})]^{-1} \mathbf{U}(\theta_{s}^{*})
\]
where
\[ \theta_0^* = B_0 \theta; \ U(\theta_0^*) = (B^T)^{-1}U(\theta); \ I(\theta_0^*) = B[I(\theta_0)]^{-1}B^T \]

### 5.1. Application to oesophageal cancer in Chinese families

This application involves the study of oesophageal cancer in 2951 nuclear families collected in Yangcheng County, Shanxi Province in China (Kwagyan, 2001). In this study, we consider as a group the nuclear family unit. The outcome variable is whether an individual is affected with oesophageal cancer or not. The objective of the study is to assess the presence and aggregation of oesophageal cancer in these families. Table 1 summarizes the distribution of number of affected individuals by family size. Of the 2951 families, 1580 (53%) had no affected individuals. The combined total number of individuals from the studied families was 14310, with mean ± sd age of 48.26 ± 18.17 years. Males comprise 56.4 and 25.4% indicated drinking alcohol.

The respondent within a family has correlated outcomes which are influenced in part or wholly by the group variables as well as the variables on the individual respondent. The main objective here is to assess the presence of familial aggregation of oesophageal cancer adjusting for measured risk factors: sex, age and alcohol consumption. Here, the model for the regression analysis of disposition is parametrization as

\[ M(Z) = \lambda, \quad D(Z) = \gamma \quad \text{and} \quad W(X) = \beta_1 \ast \text{sex} + \beta_2 \ast \text{alcohol} + \beta_3 \ast \text{age} \]

In this application, the vector of parameters \( \beta = (\beta_1, \beta_2, \beta_3) \) is the set of parameters of interest. Computations are performed using the computer program we developed CORRDAT (Bonney, Kwagyan, Slater, & Silifer, 1997b), which was linked with the likelihood optimization software MULTIMAX (Bonney, Kwagyan, & Apprey, 1997a). Computations can also be accomplished in MATLAB. Table 2 gives estimates of the correlation matrices. The correlations between the original parameters are quite high compared to those of the orthogonal parameters. In particular, the correlation between \( \lambda \) and \( \gamma \) is \(-0.179\); the correlation between \( \lambda \) and \( \beta_1 \) is \(-0.414\); and that between \( \lambda \) and \( \beta_2 \) is high, \(-0.841\). The correlations between the orthogonal parameters are near zero or nonexistent. The correlation between \( \lambda^* \) and \( \gamma^* \) is \(-0.026\); the correlation between \( \lambda^* \) and \( \beta_1^* \) is now low, 0.004 and that between \( \lambda^* \) and \( \beta_2^* \) is also low, 0.0188.

As expected, the orthogonal likelihood converged in fewer iterations than the non-orthogonal one—the orthogonal likelihood converged after 17 iterations; the non-orthogonal one converged

<table>
<thead>
<tr>
<th>Family size (s)</th>
<th>Number of Affecteds (a)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>435</td>
<td>151</td>
<td>34</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>623</td>
</tr>
<tr>
<td>4</td>
<td>536</td>
<td>215</td>
<td>52</td>
<td>16</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>819</td>
</tr>
<tr>
<td>5</td>
<td>335</td>
<td>203</td>
<td>81</td>
<td>34</td>
<td>8</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td>659</td>
</tr>
<tr>
<td>6</td>
<td>159</td>
<td>155</td>
<td>59</td>
<td>27</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td></td>
<td></td>
<td>412</td>
</tr>
<tr>
<td>7</td>
<td>74</td>
<td>95</td>
<td>46</td>
<td>12</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>232</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
<td>54</td>
<td>25</td>
<td>11</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>129</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>21</td>
<td>10</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>43</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>1580</td>
<td>906</td>
<td>314</td>
<td>110</td>
<td>24</td>
<td>13</td>
<td>3</td>
<td>1</td>
<td>2951</td>
<td></td>
</tr>
</tbody>
</table>
after 56 iterations. Estimates of the parameters are summarized in Table 3. The maximum likelihood estimate of the relative dependence parameter, $\hat{\gamma} = 0.577$ with a corresponding asymptotic 95% confidence interval of $(0.529, 0.625)$ suggests that there is significant familial aggregation of oesophageal cancer in the families sampled. Sex and age have a positive significance, while alcohol was negative. The results suggest that males are at a higher risk of getting oesophageal cancer than females and also that it is more prevalent in older people. The negative effect of alcohol seems to suggest that it has the propensity to lower the risk of oesophageal cancer in the Chinese population studied, perhaps if drank in moderation. In summary, we conclude that oesophageal cancer aggregates in the families sampled.

6. Conclusions
Parameter orthogonalization is used as an aid in computation, approximation, interpretation and elimination or reduction of the effect of nuisance parameters. The resulting algorithm that we have proposed based on the Gram–Schmidt orthogonalization process is computationally feasible. Unlike the approach of Cox–Reid which requires solving a system of differential equations to obtain orthogonality, the proposed method only requires solving a system of linear equations. Our approach has some similarities with the conjugate gradient algorithm, but is distinctly different from it. The conjugate gradient method is an optimization routine, whereas the method we have proposed is principally an approximate orthogonalization technique combined with a maximization algorithm to aid in parameter estimation. Amari (1985) claims that global orthogonalization is possible only in a special case. Our method allows for global orthogonalization, if the information matrix can be found or estimated. The approach we have proposed is based on exact calculations. The transformation is surjective, that is the original parameters can be sufficiently obtained. Clearly, the process can be cumbersome if a large number of parameters must be accurately estimated because of the inversion of the information matrix at every stage to obtain the covariance matrix and subsequently the

### Table 2. Correlation matrix from original and orthogonal parametrization. Fisher scoring algorithm used for original parametrization and Gram–Schmidt–Fisher scoring algorithm used of orthogonal parametrization

<table>
<thead>
<tr>
<th></th>
<th>Original parametrization</th>
<th>Orthogonal parametrization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\lambda$</td>
<td>$\gamma$</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>1</td>
<td>-0.179</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>-0.414</td>
<td>-0.061</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.086</td>
<td>-0.36</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.841</td>
<td>0.075</td>
</tr>
</tbody>
</table>

### Table 3. Regression analysis of disposition to oesophageal cancer. Fisher scoring algorithm used for original parametrization and Gram–Schmidt–Fisher scoring algorithm for orthogonal parametrization

<table>
<thead>
<tr>
<th></th>
<th>Estimates (SE)</th>
<th>Parameter</th>
<th>Estimates (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>-4.262 (0.103)</td>
<td>$\lambda^*$</td>
<td>-1.873 (0.028)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.577 (0.024)</td>
<td>$\gamma^*$</td>
<td>-0.008 (0.038)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.911 (0.056)</td>
<td>$\beta_1^*$</td>
<td>0.915 (0.054)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-1.131 (0.166)</td>
<td>$\beta_2^*$</td>
<td>-1.128 (0.171)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.039 (0.002)</td>
<td>$\beta_3^*$</td>
<td>0.038 (0.002)</td>
</tr>
<tr>
<td>$\log$ (Likelihood)</td>
<td>-5371.294</td>
<td>$\log$ (Likelihood)</td>
<td>-5377.876</td>
</tr>
<tr>
<td>No. of iterations</td>
<td>56</td>
<td>No. of Iterations</td>
<td>27</td>
</tr>
</tbody>
</table>
transformation matrix. The advantages, however, are that convergence is generally rapid in iteration times, sure and accurate. Block parametrization of parameters would be desirable if interest is in sets of parameters and where the dimension of the parameter space is large. The work we have presented in this article is the first attempt to consider the use of Gram–Schmidt process for estimation of the parameters. It is possible, however, that closer scrutiny, practical considerations, numerical studies for numerical stability and properties, and simulation studies to evaluate and compare convergence times would suggest modifications and refinements to the methods we have discussed.

In conclusion, the algorithm that we have proposed provides an exceedingly convenient method for multi-parameter statistical problems first to reduce numerical difficulties and second improve accuracy in parameter estimation. The method would be efficient for use of function minimization in both linear and non-linear maximum likelihood estimations and particularly useful for small parameter space estimation problems.

Funding
The work was supported in part from NIH/NCATS (grant number UL1TR000101 previously UL1RR031975) and NIH/NIMHHD (grant number 1G2MD007597).

Author details
John Kwagyan1
E-mail: jkwagyan@howard.edu
Victor Apprey1
E-mail: vapprey@howard.edu
George E. Bonney2
E-mail: ge_bonney@howard.edu
1 Design, Biostatistics & Population Studies & Department of Family Medicine, College of Medicine, Howard University, Washington, DC 20059, USA.

Citation information
Cite this article as: Gram–Schmidt–Fisher scoring algorithm for parameter orthogonalization in MLE, John Kwagyan, Victor Apprey & George E. Bonney, Cogent Mathematics (2016), 3: 1159847.

References