A unified approach to choice experiments

Ashish Das¹ and Rakhi Singh²

¹Indian Institute of Technology Bombay, Mumbai, India
²IITB-Monash Research Academy, IIT Bombay, Mumbai, India

25th April 2016

Indian Institute of Technology Bombay
Powai, Mumbai-400 076, India
A unified approach to choice experiments

ASHISH DAS
Indian Institute of Technology Bombay, Mumbai, India

RAKHI SINGH
IITB-Monash Research Academy, Mumbai, India

Abstract

Choice experiments have practical significance from the industry point of view and are useful in marketing, transport, government planning, etc. Several author-groups have contributed to the theoretical development of choice experiments and for finding optimal choice designs. The author-groups Street-Burgess and Huber-Zwerina have adopted different approaches and used seemingly different information matrices under the multinomial logit model. There have also been some confusion regarding the inference parameters expressed as linear functions of the option effects $\tau$. We discuss these aspects and highlight how these approaches are related to one another.

Recently, Sun and Dean (2016) have advocated an information matrix for orthonormal contrasts $BO\tau$, adopting a linear model approach, which is different from the standard information matrix used in the literature for choice experiments under the multinomial logit model. We study their information matrix vis-à-vis the traditional information matrix and find some lacunae in their approach. We provide an alternate linear model approach to model choice experiments such that the information matrix of $BO\tau$ is same as the information matrix under the multinomial logit model.

Keywords: Choice set; Information Matrix; Linear Model; Multinomial logit model;

1 Introduction

Choice experiments are widely used in various areas including marketing, transport, environmental resource economics and public welfare analysis. A choice experiment consists of $N$ choice sets, each containing $m$ options. A respondent is shown each of the choice sets and is in turn asked for the preferred option as per his perceived utility. The respondents choose one of the options from each choice set. Each option in a choice set is described by a set of $k$ attributes, where $i$th attribute has $v_i(\geq 2)$ levels denoted by $0, \ldots, v_i - 1$. It is assumed that there are no repeated options in a choice set. Furthermore, since the $i$th attribute has $v_i$ levels, there are a total of $L = \prod_{i=1}^{k}(v_i - 1)$ possible options. A typical option is denoted by $t_w, w = 1, \ldots, L$. A choice design is a collection of choice sets employed in a choice experiment. Though choice designs may
contain repeated choice sets, one may prefer that no two choice sets are repeated. For an excellent review of designs for choice experiments, one may refer to Street and Burgess (2012) and Großmann and Schwabe (2015).

We will denote a choice set by \( T_n \), \( 1 \leq n \leq N \), with options \((t_{n1}, t_{n2}, \ldots, t_{nm})\). Given that the options are described by the levels of the attributes, each option \( t_{nj} \) is a \( k \)-tuple \( t_{nj}^{(1)} t_{nj}^{(2)} \ldots t_{nj}^{(k)} \). For the \( j \)th option of \( N \) choice sets, let the \( N \times k \) matrix \( A_j \) represent the levels of \( k \) attributes.

In the literature, primarily, choice experiments have been discussed under the multinomial logit model (MNL model) setup. In a choice experiment, it is assumed that each respondent chooses the option having maximum utility among the other options in a choice set. Thus respondent \( \alpha \) assigns some utility \( U_{j\alpha} \) to the \( j \)th option in a choice set, \( j = 1, \ldots, m \). The respondent chooses the \( j \)th option in a choice set if \( U_{j\alpha} \geq U_{j'\alpha} \), \( j' \) being the other options in the choice set. Since these options are described by levels of the attributes, the systematic component of the utility that can be captured is denoted by \( \tau_{ja} \) and that allows us to drop the subscript \( \alpha \). Then the probability of choosing the \( j \)th option from a choice set (see, Street and Burgess (2007)) is given by,

\[
P_{nj} = P(j \text{th option is being selected in } n \text{th choice set}) = \frac{e^{\tau j}}{\sum_{j=1}^{m} e^{\tau j}}. \tag{1}
\]

Street and Burgess (2007), using the approach of El-Helbawy and Bradley (1978), gives the information matrix for estimating \( p \) parameters of interest \( B_{O\alpha} \tau \) where the \( p \) parameters are certain linear combinations of the utility vector \( \tau = (\tau_1, \ldots, \tau_L)^T \). The information matrix for \( B_{O\alpha} \tau \), as obtained by them, is \( B_{O\alpha} \Lambda B_{O\alpha}^T \), where \( B_{O\alpha} \) is the \( p \times L \) orthonormal contrast matrix corresponding to the \( p \) parameters of interest and \( \Lambda \) is the information matrix corresponding to \( \tau \). Henceforth, we will refer to this approach of obtaining the information matrix as Street-Burgess approach. As in Street and Burgess (2007), for a choice design with \( N \) choice sets and \( m \) options, under the MNL model, the information matrix for \( \tau \) is \( \Lambda = (\Lambda_{(r,r')}) \), where

\[
\Lambda_{(r,r')} = \begin{cases} 
-\sum_{n \in T^{rr'}} w_n \frac{e^{\tau r} e^{\tau r'}}{(\sum_{l \in T_n} e^{\tau l})^2}, & r \neq r', r, r' = 1, \ldots, L, \\
-\sum_{n \in T^r} w_n \frac{e^{\tau r} (\sum_{l \in T_n} e^{\tau l})}{(\sum_{l \in T_n} e^{\tau l})^2}, & r = r', r = 1, \ldots, L, 
\end{cases} \tag{2}
\]

with \( r \) and \( r' \) being the labels of the corresponding options, \( T^{rr'} \) being the set of indices of choice sets consisting of both \( \tau_r \) and \( \tau_{r'} \), \( T^r \) being the set of indices of choice sets consisting of \( \tau_r \), \( w_n = (1/N) \) if \( n \)th choice set is in the design and 0 otherwise. Under the utility-neutral MNL model assumption of all options being equally attractive, \( \Lambda_{(r,r')} \) reduces to

\[
m^2 N \Lambda_{(r,r')} = \begin{cases} 
(m - 1) a_r, & r = r', \\
-a_{r,r'}, & r \neq r'.
\end{cases} \tag{3}
\]
with \( r \) and \( r' \) the labels of the corresponding options, \( a_r \) the number of times option label \( r \) appears in the choice design and \( a_{r,r'} \) the number of times option labels \( r \) and \( r' \) occur together in choice sets of the design. In what follows, unless otherwise stated, \( \Lambda \) will refer to \( \Lambda \) as in (2). As an example, for main effects, the orthonormal contrast matrix \( B_O \) for \( p = \sum_{i=1}^{k} (v_i - 1) \) is,

\[
B_O = \begin{pmatrix}
\frac{1}{\sqrt{2}} v_1 & \frac{1}{\sqrt{2}} v_2 & \cdots & \frac{1}{\sqrt{2}} v_k \\
\frac{1}{\sqrt{2}} 1_{v_1} & \frac{1}{\sqrt{2}} B_o^{(2)} & \cdots & \frac{1}{\sqrt{2}} B_o^{(k)} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{1}{\sqrt{2}} 1_{v_1} & \frac{1}{\sqrt{2}} 1_{v_2} & \cdots & \frac{1}{\sqrt{2}} 1_{v_k}
\end{pmatrix},
\]

(4)

where \( B_o^{(i)} \) is a \((v_i - 1) \times v_i\) orthonormal contrast matrix for \( i \)th attribute at \( v_i \) levels, that is, \( B_o^{(i)T} B_o^{(i)} = I_{v_i-1} \) and \( B_o^{(i)} 1_{v_i} = 0 \). Here, \( 1_t \) is a \( t \times 1 \) vector of all ones, \( I_t \) is an identity matrix of order \( t \), and \( \otimes \) denotes the Kronecker product. For an attribute, the \( v_i \) columns of \( B_o^{(i)} \) represent an orthonormal coding for the respective levels \( 0, \ldots, (v_i - 1) \). Similarly, the columns of \( B_O \) correspond to the \( L \) options arranged in lexicographic order.

In the marketing literature, under the MNL model, usually a different approach of Huber and Zwerina (1996), following the seminal work of McFadden (1974), is followed. The utilities in this approach are modelled as

\[
U = \sum_{j=1}^{m} \beta_j\, P_{nj}.
\]

The information matrix looks superficially different and the levels of the factors are coded differently. The information matrix for \( p \) parameters of interest \( \beta_H \), as given in Huber and Zwerina (1996), is

\[
I(\beta_H) = \frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{m} (h_{nj} - \sum_{j=1}^{m} h_{nj} P_{nj})^T P_{nj} (h_{nj} - \sum_{j=1}^{m} h_{nj} P_{nj}).
\]

(5)

Under the utility-neutral MNL model, the information matrix reduces to

\[
\frac{1}{N m} \sum_{n=1}^{N} \sum_{j=1}^{m} (h_{nj} - \hat{h}_n)^T (h_{nj} - \hat{h}_n),
\]

where \( \hat{h}_n = (1/m) \sum_{j=1}^{m} h_{nj} \). It is easy to see that this information matrix for \( \beta_H \) is the sum of \( m (m - 1)/2 \) matrices corresponding to each of the pairs, that is,

\[
\frac{1}{N m} \sum_{j=1}^{m} \sum_{j'=j+1}^{m} X_{H(jj')^T} X_{H(jj')},
\]

where

\[
X_{H(jj')} = H_j - H_{j'},
\]

(6)

is the \( N \times p \) difference matrix for the \( j \)th and \( j' \)th options and the \( N \times p \) matrix \( H_j = (h_{ij}^T \ h_{ij}^T \cdots \ h_{ij}^T)^T, j = 1, \ldots, m \) is a general coded matrix corresponding to the \( A_j \).
Henceforth, we will refer to this approach of obtaining the information matrix as Huber-Zwerina approach.

The coding that is generally used in the marketing literature for describing attributes is more formally known as effects coding (see, Großmann and Schwabe (2015)). Let \( E_j \) denotes the effects coded matrix corresponding to \( A_j \), \( j = 1, \ldots, m \). Each row of \( E_j \) embeds the effects coded row vectors corresponding to the \( k \) attributes, \( j = 1, \ldots, m \). As an example, for main effects, the effects coding for level \( l \) is represented by a unit vector with 1 in the \((l+1)\)th position for \( l = 0, \ldots, v_i - 2 \), and the effects coding for level \( v_i - 1 \) is represented by \(-1\) in each of the \( v_i - 1 \) positions, \( i = 1, \ldots, k \). For example, for \( v_i = 3 \) and for main effects, effects coded vectors for \( l = 0, 1, 2 \) are \((1 0)\), \((0 1)\) and \((-1 -1)\), respectively. In what follows, under effects coding, \( H_j = E_j \). From (6), we denote the \( N \times p \) effects coded difference matrix as

\[
X_{E(jj')} = E_j - E_{j'}.
\]

Therefore, under effects coding utility-neutral MNL model, the information matrix for \( \beta_E \) is

\[
\frac{1}{N m^2} \sum_{j=1}^{m} \sum_{j'=i+1}^{m} X_{E(jj')}^T X_{E(jj')}.
\]

Alongside the MNL model, Großmann et al. (2002), Graßhoff et al. (2003) and Graßhoff et al. (2004) have simultaneously studied linear paired comparison designs analyzed under the linear paired comparison model. Unlike the MNL model, the linear paired comparison designs are applicable only for paired comparisons, that is, they are restricted to choice sets of size \( m = 2 \). The information matrix under the linear paired comparison model is proportional to the information matrix under the utility neutral MNL model for paired choice designs and hence it follows that the paired choice designs optimal under the one model are also optimal under the other model (see, Großmann and Schwabe (2015)).

The author-groups Street-Burgess and Huber-Zwerina have used seemingly different information matrices under the MNL model. There have also been some confusion regarding the inference parameters expressed as linear functions of option effects. We discuss these aspects and highlight how these approaches are related to one another.

Recently, Sun and Dean (2016) have advocated an information matrix for \( \Phi \), adopting a linear model approach, which is different from the standard information matrix used in the literature for choice experiments under the MNL model. We study their information matrix vis-à-vis the traditional information matrix and find some lacunae in their approach. We provide an alternate linear model approach to model choice experiments such that the information matrix of \( \Phi \) is same as the information matrix under the MNL model.

2 Equivalence of Huber-Zwerina and Street-Burgess approach

The mathematical derivations used while obtaining the information matrices under the Street-Burgess approach and under the Huber-Zwerina approach appears to be somewhat
different. In fact, Rose and Bliemer (2014) mentions,

“This difference in the mathematical derivations of the variance-covariance matrix has resulted in significant confusion within the literature, with claims that the Street and Burgess approach is unrelated to the more mainstream stated choice experimental design literature. This view has been further enhanced given the fact that the resulting matrix algebra used to generate the variance-covariance matrices under the two derivations appear to be very different.’

To summarize, the major differences in the two approaches are: (a) the expressions for the information matrices under the two approaches appear to be different, and (b) coding of levels in the two approaches are different.

Corresponding to $A_j$, we define an $N \times p$ orthonormally coded matrix $O_j$ where the $n$th row of $O_j$ corresponds to the $j$th option $t_{nj} = t_w$ (say) in the $n$th choice set and is equal to the $w$th row of $B^T_O$, $w = 1, \ldots, L$. We note that the derivation of the information matrix in McFadden (1974) and subsequently used by Huber and Zwerina (1996) is based on a general coding. Therefore, in particular, it holds for the orthonormal coding as in Street and Burgess (2007), that is, for $\beta_O$ with $H_j = O_j$.

We now show how for any coding structure, the two seemingly different structures of the information matrices are related. Let $B_H$ be the $p \times L$ coded matrix corresponding to the coding structure as in $H_j$. Then for $j$th option, the $n$th row of the matrix $H_j$ corresponding to an option $t_{nj} = t_w$, is same as the $w$th row of $B^T_H$. We now provide the following result, the proof of which is in the Appendix.

**Theorem 2.1.** For any coded matrix $B_H$, the information matrix for $\beta_H$ under the MNL model $\frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{m} (h_{nj} - \sum_{j=1}^{m} h_{nj} P_{nj})^T P_{nj} (h_{nj} - \sum_{j=1}^{m} h_{nj} P_{nj}) = B_H \Lambda B^T_H$, where $\Lambda$ is as defined in (2).

The following result is a Corollary to Theorem 2.1 under the utility-neutral case, that is, by assuming $P_{nj} = 1/m$ for all $n$ and $j$.

**Corollary 2.1.** For any coded matrix $B_H$, the information matrix for $\beta_H$ under the utility-neutral MNL model, $\frac{1}{m^2 N} \sum_{j=1}^{m} \sum_{j' = j+1}^{m} X^T_{H(jj')} X_{H(jj')} = B_H \Lambda B^T_H$, where $\Lambda$ is as defined in (3).

Therefore, we have shown that once the codings is fixed, the two expressions which appear different, are, in fact, equal. This also implies that whether one differentiates with respect to $\tau$ first or directly with respect to $\beta$, one gets the same information matrix.

Using (6), we define an $N \times p$ matrix

$$X_{O(jj')} = O_j - O_{j'},$$

as the corresponding orthonormally coded difference matrix. Now for effects coding, we can define a $p \times L$ effects coded matrix $B_E$, where $w$th row of $B^T_E$ is the effects coding for
\( t_n \) as in \( E_j \) in (7) or equivalently as in Großmann and Schwabe (2015). As an example, for estimation of the main effects, \( B_E \) for \( p = \sum_{i=1}^{k} (v_i - 1) \) is

\[
B_E = \begin{pmatrix}
B_e^{(1)} \otimes 1_{m_2}^T \otimes \cdots \otimes 1_{m_k}^T \\
1_{v_1} \otimes B_e^{(2)} \otimes \cdots \otimes 1_{m_k}^T \\
\vdots \otimes \cdots \otimes \cdots \otimes \cdots \\
1_{v_1} \otimes 1_{v_2} \otimes \cdots \otimes B_e^{(k)}
\end{pmatrix},
\]

where \( B_e^{(i)} \) is a \((v_i - 1) \times v_i\) effects coded matrix for \( i \)th attribute at \( v_i \) levels and for obtaining effects coded \( B_e^{(i)} \), effects coding corresponding to level \( l \) is put as the \( l \)th column of \( B_e^{(i)} \). For example, for \( v_i = 3 \), \( B_e^{(i)} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \).

**Corollary 2.2.** The following are two special cases of Corollary 2.1 under the utility-neutral MNL model.

(i) For orthonormally coded \( B_O \) as in Street and Burgess (2007), the information matrix \( \mathcal{I}(\beta_O) = \frac{1}{mN} \sum_{j=1}^{m} \sum_{j'=i+j}^{j+i} X_{O(j')j}^T X_{O(j')} = B_O \Lambda B_O^T = \mathcal{I}(B_O \tau) \). In other words, the variance \( V(\hat{\beta}_O) = (\frac{1}{mN} \sum_{j=1}^{m} \sum_{j'=i+j}^{j+i} X_{E(j')j}^T X_{E(j')})^{-1} = (B_O \Lambda B_O^T)^{-1} = V(\hat{\tau}) \).

(ii) For effects coded \( B_E \), the information matrix \( \mathcal{I}(\beta_E) = \frac{1}{mN} \sum_{j=1}^{m} \sum_{j'=i+j}^{j+i} X_{E(j')j}^T X_{E(j')} = B_E \Lambda B_E^T \). In other words, the variance is \( V(\hat{\beta}_E) = (\frac{1}{mN} \sum_{j=1}^{m} \sum_{j'=i+j}^{j+i} X_{E(j')j}^T X_{E(j')})^{-1} = (B_E \Lambda B_E^T)^{-1} \).

**Remark 2.1.** We note that in Corollary 2.2 (i), we have \( V(\hat{\beta}_O) = V(\hat{\beta}_E) \). However, in Corollary 2.2 (ii), we mention the variance only of \( \hat{\beta}_E \), not mentioning anything in terms of \( \hat{\tau} \). We elaborate on the same in the remainder of this section.

Page 77–78, Chapter 3 of Street and Burgess (2007) gives an impression (through an example) that if \( B_E \) is the effects coding matrix, then under the utility-neutral setup, \( \mathcal{I}(B_E \tau) = B_E \Lambda B_E^T = \frac{1}{mN} \sum_{j=1}^{m} \sum_{j'=i+j}^{j+i} X_{E(j')j}^T X_{E(j')} \). In contrast, Großmann and Schwabe (2015) have indicated that under the utility-neutral setup, for an optimal paired choice design \( d^* \), \( \mathcal{I}_d(\beta_E) = \mathcal{I}_d(SG\tau) = (S(G\Lambda G^T)^{-}S^T)^{-} = M_{E^*} \), where \( M_{E^*} = \text{diag}(M_{v_1}, M_{v_2}, \ldots, M_{v_k}) \) with \( M_{v_i} = \frac{2}{m_N-1} (I_{v_i-1} + J_{v_i-1}) \). Here, \( J_t \) is a \( t \times t \) matrix of all ones. \( G \) is obtained from (4) by replacing every \((v_i - 1) \times v_i\) matrix \( B_o^{(i)} \) with the \( v_i \times v_i \) centering matrix \( K_{v_i} = I_{v_i} - \frac{1}{v_i} J_{v_i} \) and \( S \) is a rectangular block diagonal matrix with \( k \) diagonal blocks of dimension \((v_i - 1) \times v_i\) being given by \( D_{v_i} = (\sqrt{\frac{mN}{2}} I_{v_i-1}, 0_{v_i-1 \times 1}) \).

We now study the following example of an optimal paired choice design \( d^* \).

**Example 2.1.** Consider an optimal paired choice design \( d^* \) with \( k = 2, v_1 = v_2 = 3 \) and \( N = 9 \).
\[ d^* = (00, 11), (10, 21), (20, 01), (01, 12), (11, 22), (21, 02), (02, 10), (12, 20), (22, 00) \]

Then, it is easy to see that,
\[ \mathcal{I}_{d^*}(\beta_E) = \mathcal{I}_{d^*}(SG\tau) = \left( S(G\Lambda G^T)^{-1}S^T \right)^{-1} = \begin{pmatrix} 0.50 & 0.25 & 0 & 0 \\ 0.25 & 0.50 & 0 & 0 \\ 0 & 0 & 0.50 & 0.25 \\ 0 & 0 & 0.25 & 0.50 \end{pmatrix} = M_{\xi^*}. \]

Also,
\[ \mathcal{I}_{d^*}(B_E\tau) = B_E\Lambda B_E^T = \begin{pmatrix} 0.50 & 0.25 & 0 & 0 \\ 0.25 & 0.50 & 0 & 0 \\ 0 & 0 & 0.50 & 0.25 \\ 0 & 0 & 0.25 & 0.50 \end{pmatrix} = M_{\xi^*}. \]

However, since
\[ B_E = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 \\ 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 \end{pmatrix} \]

and
\[ SG = \begin{pmatrix} 0.22 & 0.22 & 0.22 & -0.11 & -0.11 & -0.11 & -0.11 & -0.11 \\ -0.11 & -0.11 & -0.11 & 0.22 & 0.22 & -0.11 & -0.11 & -0.11 \\ 0.22 & -0.11 & -0.11 & 0.22 & -0.11 & 0.22 & -0.11 & -0.11 \\ -0.11 & 0.22 & -0.11 & -0.11 & 0.22 & -0.11 & 0.22 & -0.11 \end{pmatrix}, \]

it follows that though \( \mathcal{I}_{d^*}(B_E\tau) = \mathcal{I}_{d^*}(SG\tau) \), still \( B_E \neq SG \). Thus, there exists a lack of clarity on what \( \beta_E \) is in terms of \( \tau \). Is it that \( \beta_E = B_E\tau \) or \( \beta_E = SG\tau \) or is it something else?

In what follows, we try to address these ambiguities under a more general unrestricted design setup on any design \( d \) having \( m \geq 2 \) and with no restriction of utility-neutrality.

**Theorem 2.2.** For a general coding \( B_H \),

(i) \( \text{Var}(B_H\hat{\tau}) = (B_HB_H^T)(B_H\Lambda B_H^T)^{-1}(B_HB_H^T) \), and 

(ii) \( \text{Var}((B_HB_H^T)^{-1}B_H\hat{\tau}) = (B_H\Lambda B_H^T)^{-1} = V(\hat{\beta}_H) \).

**Proof.** For every \( p \times L \) matrix \( B_H \) whose rows are not necessarily orthogonal but that span the same vector space as the rows of \( p \times L \) matrix \( B_O \), there exists a non-singular matrix \( Q \) of order \( p \) such that
\[ B_H = QB_O. \quad (9) \]

Now, \( B_H = QB_O \) implies \( B_O = Q^{-1}B_H \). Also, since \( B_OB_O^T = I_p \), it follows that \( B_HB_H^T = QB_OB_O^TQ^T = QQ^T \). Then, from Corollary 2.2 (i),
\[ \text{Var}(B_H \hat{\tau}) = \text{Var}(QB_O \hat{\tau}) = \text{QVar}(B_O \hat{\tau})Q^T = Q(B_O \Lambda B_O^{-1})Q^T = \{(Q^T)^{-1}(B_O \Lambda B_O^{-1})Q^{-1}\}^{-1} = \{(QQ^T)^{-1}(B_H \Lambda B_H^{-1})(B_H B_H^{-1})^{-1}\}^{-1} = (B_H B_H^{-1})(B_H \Lambda B_H^{-1})^{-1}(B_H B_H^{-1}). \] (10)

Also, from (10) and Theorem 2.1, it follows that,

\[ \text{Var}(B_H B_H^{-1} B_H \hat{\tau}) = (B_H B_H^{-1})^{-1} \text{Var}(B_H \hat{\tau})(B_H B_H^{-1})^{-1} = (B_H \Lambda B_H^{-1})^{-1} = V(\hat{\beta}_H). \]

We now have the following Corollary as special cases when (i) \( B_H = B_O \) and (ii) \( B_H = B_E \).

**Corollary 2.3.** For \( B_H = B_O \) and \( B_H = B_E \), the following holds.

(i) \( \text{Var}(B_O \hat{\tau}) = (B_O \Lambda B_O^{-1}) = \text{Var}(\hat{\beta}_O); \) and \( \mathcal{I}(B_O \tau) = B_O \Lambda B_O^{-1}. \)

(ii) \( \text{Var}(B_E \hat{\tau}) = (B_E \Lambda B_E^{-1})(B_E B_E^{-1})^{-1}(B_E \Lambda B_E^{-1})^{-1} = \mathcal{I}(B_E \tau) = (B_E B_E^{-1})(B_E B_E^{-1})^{-1}. \)

(iii) \( \text{Var}(B_E B_E^{-1} B_E \hat{\tau}) = (B_E B_E^{-1})^{-1} = \text{Var}(\hat{\beta}_E); \) and \( \mathcal{I}(B_E B_E^{-1} B_E \tau) = B_E \Lambda B_E^{-1}. \)

This establishes the correct information matrix of \( B_E \tau \) as against the impression created in Street and Burgess (2007). Following Großmann and Schwabe (2015), for a balanced design \( d^* \), the information matrix is \( \mathcal{I}_{d^*}(\beta_E) = \mathcal{I}_{d^*}(SG \tau) = M_{d^*} \) while it follows from Corollary 2.2 (ii) and Corollary 2.3 (iii) that \( \mathcal{I}_{d^*}(B_E B_E^{-1} B_E \tau) = M_{d^*} \). We now show that both results are equivalent.

**Theorem 2.3.** \( (B_E B_E^{-1})^{-1} B_E = SG. \)

**Proof.** From (4), \( G \) can be written as,

\[
G = \begin{pmatrix}
K_{v_1} \otimes \frac{1}{\sqrt{v_2}} 1_{v_2}^T \otimes \cdots \otimes \frac{1}{\sqrt{v_k}} 1_{v_k}^T \\
\frac{1}{\sqrt{v_2}} 1_{v_1}^T \otimes K_{v_2} \otimes \cdots \otimes \frac{1}{\sqrt{v_k}} 1_{v_k}^T \\
\vdots \\
\frac{1}{\sqrt{v_1}} 1_{v_1}^T \otimes \frac{1}{\sqrt{v_2}} 1_{v_2}^T \otimes \cdots \otimes K_{v_k}
\end{pmatrix}
\]

8
Also, from definition, $S$ is a rectangular block diagonal matrix given by,

$$S = \frac{1}{\sqrt{L}} \begin{pmatrix}
D_{v_1} & 0_{(v_1-1)\times v_2} & \ldots & 0_{(v_1-1)\times v_k} \\
0_{(v_2-1)\times v_1} & D_{v_2} & \ldots & 0_{(v_2-1)\times v_k} \\
\vdots & \vdots & \ddots & \vdots \\
0_{(v_k-1)\times v_1} & 0_{(v_k-1)\times v_2} & \ldots & D_{v_k}
\end{pmatrix}.$$  

Therefore, on multiplication,

$$SG = \frac{1}{L} \begin{pmatrix}
T_{v_1} \otimes 1_{v_2}^T \otimes \ldots \otimes 1_{v_k}^T \\
1_{v_1}^T \otimes T_{v_2} \otimes \ldots \otimes 1_{v_k}^T \\
\vdots \vdots \vdots \vdots \\
1_{v_1}^T \otimes 1_{v_2}^T \otimes \ldots \otimes T_{v_k}
\end{pmatrix}, \quad (11)$$

where $T_{v_i}$ is the $(v_i - 1) \times v_i$ matrix of the first $v_i - 1$ rows of $v_iK_{v_i}$. Now, it is easy to see that

$$(B_EB_E^T)^{-1} = \frac{1}{L} \begin{pmatrix}
v_1 I_{v_1-1} - J_{v_1-1} & 0 & \ldots & 0 \\
0 & v_2 I_{v_2-1} - J_{v_2-1} & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & v_k I_{v_k-1} - J_{v_k-1}
\end{pmatrix},$$

and that $(v_i I_{v_i-1} - J_{v_i-1})B_E^{(i)} = T_{v_i}$. Therefore,

$$(B_EB_E^T)^{-1}B_E = \frac{1}{L} \begin{pmatrix}
T_{v_1} \otimes 1_{v_2}^T \otimes \ldots \otimes 1_{v_k}^T \\
1_{v_1}^T \otimes T_{v_2} \otimes \ldots \otimes 1_{v_k}^T \\
\vdots \vdots \vdots \vdots \\
1_{v_1}^T \otimes 1_{v_2}^T \otimes \ldots \otimes T_{v_k}
\end{pmatrix} = SG. \quad (12)$$

Then the result follows from (11) and (12).

Graßhoff et al. (2004) characterized optimal designs with $m = 2$ under effects coding utility-neutral MNL model for estimating the main effects. In what follows, we now extend their results for $m > 2$.

**Theorem 2.4.** For an optimal design $d^*$ for estimating the main effects corresponding to $k$ attributes each at $v$ levels, that is, where $I_{d^*}(B_O\hat{\tau}) = B_O\Lambda B_O^T = \alpha I_{k(v-1)}$ with $\alpha$ as in Theorem 6.3.1 of Street and Burgess (2007), the following are true for the information matrices of $d^*$ for inferring on $\beta_E$ and $B_E\tau$.

1. $I_{d^*}(\beta_E) = I_{d^*}((B_EB_E^T)^{-1}B_E\tau) = \alpha B_EB_E^T = \alpha v^{k-1}(I_k \otimes (I_{v-1} + J_{v-1}))$.
2. $I_{d^*}(B_E\tau) = \alpha(B_EB_E^T)^{-1} = (\alpha / v^{k-1})(I_k \otimes (I_{v-1} - (1/v)J_{v-1})).$
Proof. Using (9) with $B_H = B_E$, we have

(i) $\mathcal{I}_d(\beta_E) = \mathcal{I}_d((B_EB_E^T)^{-1}B_E) = B_EAB_E^T = QB_O\Lambda B_O^T Q^T = Q(\alpha I_{k(v-1)})Q^T = \alpha QQ^T = \alpha B_EB_E^T = \alpha v^{-1}(I_k \otimes (I_{v-1} + J_{v-1})).$

Furthermore,

(ii) $\mathcal{I}_d(B_E\tau) = (B_EB_E^T)^{-1}(B_E\Lambda B_E^T)(B_EB_E^T)^{-1} = (B_EB_E^T)^{-1}(\alpha B_EB_E^T)(B_EB_E^T)^{-1} = \alpha(B_EB_E^T)^{-1} = (\alpha v^{-1})(I_k \otimes (I_{v-1} - (1/v)J_{v-1})).$

\[ \blacksquare \]

3 Under the MNL model is $V(B_O\hat{\tau}) = B_O\Lambda^{-1}B_O^T$?

Recently, Sun and Dean (2016) have obtained locally $A$-optimal choice designs. They concluded that for orthonormal $B_O$, an alternate form of the variance of $B_O\tau$ is, $V(B_O\hat{\tau}) = B_O\Lambda^{-1}B_O^T$. A similar statement was made earlier by Großmann and Schwabe (2015) where they said that for any non-orthogonal $B_H$, under the MNL model, $\mathcal{I}(B_H\tau) = (B_H\Lambda^{-1}B_H^T)^{-1}$ or that $V(B_H\hat{\tau}) = B_H\Lambda^{-1}B_H^T$. The results of Sun and Dean (2016) relies on a linear model setup that they have adopted.

A choice design is said to be connected if for estimating $\beta_H$, the information matrix $B_H\Lambda B_H^T$ is non-singular. Let $\mathcal{D}(k, m, N)$ be the class of connected choice designs with $m$ options, $N$ choice sets having $k$ attributes with the $i$th attribute at $v_i$ levels, $i = 1, \ldots, k$, for estimating $\beta_O = B_O\tau$. Also, in line with Sun and Dean (2016), let $\mathcal{D}_B(k, m, N)$ be the subclass of $\mathcal{D}(k, m, N)$ in which all profiles are present and in which $B_O\tau$ is estimable (that is, rows of $B_O$ belong to row space of $\Lambda$).

Note that under the MNL model, there is no ambiguity to the fact that $V(B_O\hat{\tau}) = (B_O\Lambda B_O^T)^{-1}$. Therefore, a question that arise is whether one can generally consider $V(B_O\hat{\tau}) = B_O\Lambda^{-1}B_O^T$? Also, if $V(B_O\hat{\tau}) = B_O\Lambda^{-1}B_O^T$ then is it that $B_O\Lambda^{-1}B_O^T = (B_O\Lambda B_O^T)^{-1}$?

We note that $B_O\Lambda^{-1}B_O^T$ is not invariant to the choice of the generalized-inverse unless the rows of $B_O$ belong to the row-space of $\Lambda$. Equivalently, it can be shown that $B_O\Lambda^{-1}B_O^T$ is invariant to the choice of the generalized-inverse if and only if $B_O\Lambda^{-1} = B_O$. In the linear model setup, this ensures estimability of $B_O\tau$. Such a restriction is not required for estimation of $\beta_O = B_O\tau$ under the MNL model. Therefore, in general, it appears inappropriate to say that under the MNL model, the information matrix for the $p$ parameters of interest is $(B_O\Lambda^{-1}B_O^T)^{-1}$. Furthermore, though Street and Burgess (2007) have proved that $\mathcal{I}(\tau) = \Lambda$ but the use of $V(\hat{\tau}) = \Lambda^{-1}$ in choice design literature did not exist until the Großmann and Schwabe (2015) and Sun and Dean (2016).

There also arise a question as to whether $(B_O\Lambda^{-1}B_O^T) = (B_O\Lambda B_O^T)^{-1}$ even when the rows of $B_O$ belong to the row-space of $\Lambda$. Sun (2012) has shown that if $\text{rank}(B_O) = \text{rank}(\Lambda)$, then the determinant as well as the trace of the two matrices $(B_O\Lambda^{-1}B_O^T)$ and $(B_O\Lambda B_O^T)^{-1}$ become equal which in turn implies that in such cases, the $D$- and the $A$-optimal designs under both the information matrices would remain same. However, they have also noted that situations where $\text{rank}(B_O) = \text{rank}(\Lambda)$ are not very practical.

Example 3.1. Consider two 2-level paired choice designs with $k = 3$ and $N = 5$ for estimating the main effects. Example 3.3 of Sun and Dean (2016) also consider the same
parameters. There are a total of \((\binom{26}{6})\) = 98,280 designs of which 95,544 designs belong to \(\mathcal{D}(3, 2, 5)\) while only 96 designs belong to \(\mathcal{D}_B(3, 2, 5)\). Among these 96 designs, 24 designs minimize \(\text{trace}(B_O\Lambda^-B_O^T)\) and of these 24 designs, only 12 designs minimizes \(\text{trace}(B_O\Lambda B_O^T)^{-1}\). Thus we see that, although all the 24 designs are A-optimal as per Sun-Dean’s A-optimality criteria, only 12 of them are A-optimal as per Street-Burgess or Huber-Zwerina criteria. We now give two designs \(d_1\) and \(d_2\), where \(d_1\) is A-optimal under Street-Burgess setup, while \(d_1\) and \(d_2\) both are A-optimal under the Sun-Dean approach.

\[
d_1 = \begin{pmatrix} 000, & 111 \\ 001, & 110 \\ 010, & 101 \\ 011, & 100 \\ 000, & 110 \end{pmatrix} \quad \text{with} \quad B_O\Lambda^-B_O^T = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix} \quad \text{and} \quad (B_O\Lambda B_O^T)^{-1} = \begin{pmatrix} 8.333 & -1.667 & 0 \\ -1.667 & 8.333 & 0 \\ 0 & 0 & 10 \end{pmatrix}.
\]

\[
d_2 = \begin{pmatrix} 000, & 111 \\ 001, & 110 \\ 010, & 101 \\ 011, & 100 \\ 000, & 100 \end{pmatrix} \quad \text{with} \quad B_O\Lambda^-B_O^T = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix} \quad \text{and} \quad (B_O\Lambda B_O^T)^{-1} = \begin{pmatrix} 8 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix}.
\]

It is also noted that Sun and Dean (2016) have adopted an algorithm to pick the best designs based on the contribution of choice sets and, in the above example, it would not pick the designs like \(d_2\) (since the choice set contribution of the added 5th pair is less).

We now give another example where none of the A-optimal designs under the Street-Burgess approach are A-optimal under the Sun-Dean approach.

**Example 3.2.** Consider two 2-level designs with \(k = 3\) and \(N = 6\) for estimating the main effects. A total of \(\binom{28}{6}\) = 376,740 designs are possible, of which 6,298 designs belong to \(\mathcal{D}_B(3, 2, 6)\). There are only 6 designs in \(\mathcal{D}_B(3, 2, 6)\) having the minimum value of \(\text{trace}(B_O\Lambda^-B_O)\) as 30. As an A-optimal design in \(\mathcal{D}_B(3, 2, 6)\) as per the Sun-Dean approach, we provide \(d_1\) as below.

\[
d_1 = \begin{pmatrix} 000, & 111 \\ 001, & 110 \\ 010, & 101 \\ 000, & 011 \end{pmatrix} \quad \text{with} \quad B_O\Lambda^-B_O^T = \begin{pmatrix} 12 & 0 & 0 \\ 0 & 9 & -3 \\ 0 & -3 & 9 \end{pmatrix} \quad \text{and} \quad (B_O\Lambda B_O^T)^{-1} = \begin{pmatrix} 12 & 0 & 0 \\ 0 & 9 & -3 \\ 0 & -3 & 9 \end{pmatrix}.
\]

On the other hand, 373,796 designs belong to \(\mathcal{D}(3, 2, 6)\), of which 12 designs (all different from the 6 designs which minimizes \(\text{trace}(B_O\Lambda^-B_O)\)) have minimum value of \(\text{trace}(B_O\Lambda B_O^T)^{-1}\) as 28. Incidentally, all these 12 designs also belong to \(\mathcal{D}_B(3, 2, 6)\). As an A-optimal design in \(\mathcal{D}(3, 2, 6)\) as per the Street-Burgess approach, we provide \(d_2\) as below.
Therefore, we conclude that for estimating $B_O \tau$, one should use the information matrix as obtained in the Street and Burgess (2007) and not the one in Sun and Dean (2016).

Sun and Dean (2016) have adopted a linear model approach to calculate the choice set contribution exploiting the ‘linearization’ of the MNL model. Their locally linearized model is $Y_n = F_n^T \tau + \epsilon_n$, where $Y_n = (Y_{n1}, \ldots, Y_{nm})^T$ is the observation vector corresponding to the nth choice set and $F_n = (F_{n1}, \ldots, F_{nm})^T$ is a $v \times m$ matrix defined appropriately. The linear model approach of Sun and Dean (2016) estimates $\tau$ first. This inherently restricts the class of competing designs in order to ensure estimability of $B_O \tau$. Moreover, it leads to a somewhat different information matrix for estimating $B_O \tau$ which is not in line with the information matrix of $B_O \tau$ as obtained by Street-Burgess and Huber-Zwerina.

In what follows, we propose an alternate linear model approach for choice designs which corrects the defects, as highlighted above, in the Sun-Dean’s approach.

On lines similar to Sun and Dean (2016), let $Y_n = (Y_{n1}, \ldots, Y_{nm})^T$ be the vector of observations made from choice set $T_n$ and that $Y_{nj} = 1$ if the jth option is selected in the nth choice set and 0 otherwise, $j = 1, \ldots, m$. Let $z_{nj} = p_{nj}(o_{nj} - \sum_{j=1}^{m} o_{nj} p_{nj})$, where $o_{nj}$ is the nth row of $O_j$. Let $Z_n = (z_{n1}^T \ldots z_{nm}^T)^T$. We propose the linear model

$$Y_n = Z_n B_O \tau + \epsilon_n,$$  \hspace{1cm} (13)

where, following Sun and Dean (2016), $Y_{nj}$ follows multinomial distribution with $P(Y_{nj} = 1) = p_{nj}$, $\text{Cov}(Y_{nj1}, Y_{nj2}) = -p_{nj1} p_{nj2}$ for $j_1 \neq j_2 = 1, \ldots, m$ and $V(Y_{nj}) = p_{nj}(1 - p_{nj})$, $j = 1, \ldots, m$. Thus, the variance-covariance matrix of $Y_n$ is given by $\Sigma_n = \text{Cov}(Y_n) = (D_n - D_n J_m D_n)$, where $D_n = \text{diag}(p_{n1}, \ldots, p_{nm})$.

Then, the information matrix for $\tau$ for the nth choice set is, $\mathcal{I}(\tau)_n = B_O^T Z_n^T \Sigma_n^{-1} Z_n B_O$. One generalized inverse of $\Sigma_n$ is $D_n^{-1}$ and therefore,

$$\mathcal{I}(\tau)_n = B_O^T Z_n^T D_n^{-1} Z_n B_O.$$  \hspace{1cm} (14)

For a design with $N$ choice sets, the model (13) becomes

$$Y = ZB_O \tau + \epsilon,$$  \hspace{1cm} (15)

with $Y = (Y_1, \ldots, Y_N)^T$, $Z = (Z_1, \ldots, Z_N)^T$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_N)^T$. Then, using (14), the average information matrix for $\tau$ can be written as

$$\mathcal{I}(\tau) = \frac{1}{N} \sum_{n=1}^{N} (B_O^T Z_n^T D_n^{-1} Z_n B_O) = B_O^T Z^T D^{-1} Z B_O,$$  \hspace{1cm} (16)
where $D = \text{diag}(D_1, \ldots, D_n)$. Now, for the linear model as in (15), $B_O \tau$ is always estimable since 
$\text{rank}(B_O^T Z^T D^{-1} Z B_O) = \text{rank}(D^{-1/2} Z B_O) = \text{rank} \left( \frac{D^{-1/2} Z B_O}{B_O} \right)$. Also, we know that if $B_O \tau$ is estimable, then $B_O (B_O^T Z^T D^{-1} Z B_O)^{-1} (B_O^T Z^T D^{-1} Z B_O) = B_O$.

In what follows, we establish that $V(B_O \hat{\tau})$, under the linear model (15), is equal to the variance as obtained by Street and Burgess (2007).

**Theorem 3.1.** Under the linear model (15), $V(B_O \hat{\tau}) = B_O (B_O^T B_O \Lambda B_O^T B_O)^{-1} B_O^T = (B_O \Lambda B_O)^{-1},$ where $\Lambda$ is as in (2).

**Proof.** Since $B_O (B_O^T Z^T D^{-1} Z B_O)^{-1} (B_O^T Z^T D^{-1} Z B_O) = B_O$, post-multiplying both sides by $B_O^T (Z^T D^{-1} Z)^{-1}$, we get $V(B_O \hat{\tau}) = B_O (B_O^T Z^T D^{-1} Z B_O)^{-1} B_O^T = (Z^T D^{-1} Z)^{-1}$. Now, from the definition of $Z$ and from Theorem 2.1, it is easy to see that $Z^T D^{-1} Z = \frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{m} (o_{nj} - \sum_{j=1}^{m} o_{nj} p_{nj})^T p_{nj} (o_{nj} - \sum_{j=1}^{m} o_{nj} p_{nj}) = B_O \Lambda B_O$. \hfill \blacksquare

**Corollary 3.1.** Under the utility-neutral linear model (15) with $p_{nj} = (1/m)$, $V(B_O \hat{\tau}) = B_O (B_O^T B_O \Lambda B_O^T B_O)^{-1} B_O^T = (B_O \Lambda B_O)^{-1}$ where $\Lambda$ is as in (3).

### 4 Concluding Remarks

While addressing the confusions of how two seemingly different approaches for deriving the information matrices under the MNL model, we could establish that the two approaches are equivalent. Though, in a passing remark, Rose and Bliemer (2014) have indicated that they were able to show Street-Burgess designs as a special case of the designs obtained by other researchers like Huber-Zwerina, we have theoretically establish a unified approach to choice experiments. Also, for the non-orthonormal coding, we have obtained a simple parametric function of $\tau$, which is the correct function that is being inferred upon under the Huber-Zwerina approach.

The impression provided in Sun and Dean (2016) for obtaining $V(B_O \hat{\tau})$, under the MNL model, using $V(\hat{\tau}) = \Lambda^{-}$ is not appropriate. Also, the class $D_B(k, m, N)$ of Sun-Dean is very restrictive. For example, for 3 attributes at 2 levels each, the class $D_B(3, 2, 4)$ contains only 1 design whereas the class $D(3, 2, 4)$ has 18,519 designs. Moreover, the total number of designs in $D_B(3, 2, 6)$ is less than 2% of the total number of designs in $D(3, 2, 6)$ and the designs in $D_B(3, 2, 5)$ is less than 0.1% of the total number of designs in $D(3, 2, 5)$. We have shown that for estimating $B_O \tau$, we can adopt an alternate expression for $V(\hat{\tau})$ as given in (16). This is achieved through our proposed linear model (15) leading to the same variance-covariance matrix of $B_O \hat{\tau}$ used in Street and Burgess (2007) and Huber and Zwerina (1996) for identifying optimal designs.

### Acknowledgement

Ashish Das’s work is partially supported by the Department of Science and Technology Project Grant 14DST007.
References


Appendix

Proof of Theorem 2.1

It is easy to see that

\[
\Lambda = \frac{1}{N} \sum_{n=1}^{N} \Delta_n = \frac{1}{N} \sum_{n=1}^{N} \sum_{1<j'=m} \Delta_{n(j')}(r, r'),
\]

where

\[
\left( \sum_{l \in T_n} e_l^T \right)^2 \Delta_{n(j')}(r, r') = \begin{cases} 
-e^r e^{r'}, & r \neq r', r' = 1, \ldots, L, \\
+e^r e^{r'}, & r = r', r = 1, \ldots, L,
\end{cases}
\]

As per the definition, options are lexicographically arranged in \( \Lambda \) as well as \( B_H \). Also, the row in \( H_1 \) corresponding to an option \( \tau_w \), is given by \( w \)th column of \( B_H^T \). Let \( B_H \) is a matrix of order \( p \times L \).

Let \( B_H = [B_1 \ b_r \ B_2 \ b_r' \ B_3] \), where \( B_1 \) is of order \( p \times (r - 1) \), \( B_2 \) is of order \( p \times (r' - r - 1) \), and \( B_3 \) is of order \( p \times (L - r') \). Without loss of generality, let \( h_{nj} \) and \( h_{n'j'} \) correspond to the \( r \)th and the \( r' \)th lexicographic labels of \( r \)th and \( r' \)th options respectively, with \( r < r' \). Then, \( b_r = h_{nj}^T \) and \( b_r' = h_{n'j'}^T \). Here \( ^T \) denotes the transposition. Therefore,

\[
B_H \Delta_{n(j')} B_H^T = \frac{1}{N} B_H \sum_{n=1}^{N} \left( \frac{1}{m^2} \sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} \Delta_{n(j')}(r, r') \right) B_H^T
\]

\[
= \frac{1}{m^2N} \sum_{n=1}^{N} \sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} B_H \Delta_{n(j')}(r, r') B_H^T.
\]

Now, by definition,

\[
\left( \sum_{l \in T_n} e_l^T \right)^2 \Delta_{n(j')}(r, r') = e^r e^{r'} \begin{bmatrix} 0 & w_{njj'}^T & 0_{L \times (r' - r - 1)} & -w_{njj'}^T & 0_{L \times (L - r')} \end{bmatrix},
\]

where \( w_{njj'} = \begin{bmatrix} 0_{1 \times (r-1)} & 1 & 0_{1 \times (r' - r - 1)} & -1 & 0_{1 \times (L - r')} \end{bmatrix} \).

Then,

\[
B_H \Delta_{n(j')} B_H^T = \frac{e^r e^{r'}}{\left( \sum_{l \in T_n} e_l^T \right)^2} \begin{bmatrix} 0_{p \times (r-1)} & (h_{nj}^T - h_{n'j'}^T) & 0_{p \times (r' - r - 1)} & (h_{nj}^T - h_{n'j'}^T) & 0_{p \times (L - r')} \end{bmatrix} B_H^T
\]

\[
= \frac{e^r e^{r'}}{\left( \sum_{l \in T_n} e_l^T \right)^2} ((h_{nj}^T - h_{n'j'}^T) h_{n'j'} - (h_{nj}^T - h_{n'j'}^T) h_{nj}) = \frac{e^r e^{r'}}{\left( \sum_{l \in T_n} e_l^T \right)^2} (h_{nj}^T - h_{n'j'}^T) (h_{n'j'} - h_{nj}).
\]

From the definition, we get,

\[
B_H \Delta_{n(j')} B_H^T = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{\left( \sum_{l \in T_n} e_l^T \right)^2} \sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} e^r e^{r'} (h_{nj} - h_{n'j'})^T (h_{nj} - h_{n'j'})
\]

\[
= \frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} P_{nj} P_{n'j'} (h_{nj} - h_{n'j'})^T (h_{nj} - h_{n'j'}).\]
Upon simple rearrangement and using the fact that \( \sum_{j=1}^m P_{nj} = 1 \) for each \( n = 1, \ldots, N \), we get,

\[
B_H \Lambda B_H^T = \frac{1}{N} \sum_{n=1}^N \left( \sum_{j=1}^m h_{nj}^T h_{nj} P_{nj}(1 - P_{nj}) - \sum_{j \neq j'}^m h_{nj}^T h_{nj'} P_{nj} P_{nj'} \right)
\]

\[
= \frac{1}{N} \sum_{n=1}^N \left( \sum_{j=1}^m h_{nj}^T h_{nj}(P_{nj} - 2P_{nj}^2) + \sum_{j=1}^m h_{nj}^T h_{nj} P_{nj}^2 \sum_{j=1}^m P_{nj} - 2 \sum_{j \neq j'}^m h_{nj}^T h_{nj'} P_{nj} P_{nj'} \right)
\]

\[
+ \sum_{j \neq j'}^m h_{nj}^T h_{nj'} P_{nj} P_{nj'} \sum_{j=1}^m P_{nj} \right)
\]

\[
= \frac{1}{N} \sum_{n=1}^N \sum_{j=1}^m \sum_{l=1}^m \left( h_{nj}^T h_{nj} - \sum_{l=1}^m h_{nl}^T h_{nj} P_{nl} - \sum_{l=1}^m h_{nj}^T h_{nl} P_{nl} + \sum_{l_1 \neq l_2} \sum_{l=1}^m h_{nl}^T h_{nl} P_{nl} \right)
\]

\[
= \frac{1}{N} \sum_{n=1}^N \sum_{j=1}^m \left( h_{nj}^T P_{nj} - \sqrt{P_{nj}} \sum_{j=1}^m h_{nj} P_{nj} \right)^T \left( h_{nj}^T P_{nj} - \sqrt{P_{nj}} \sum_{j=1}^m h_{nj} P_{nj} \right)
\]

Hence,

\[
B_H \Lambda B_H^T = \frac{1}{N} \sum_{n=1}^N \sum_{j=1}^m \left( h_{nj}^T P_{nj} - \sqrt{P_{nj}} \sum_{j=1}^m h_{nj} P_{nj} \right) P_{nj} \left( h_{nj}^T P_{nj} - \sqrt{P_{nj}} \sum_{j=1}^m h_{nj} P_{nj} \right)
\].