The Bouchard Prize: Two Tasks

The following outlines the two tasks that must be completed to win the Bouchard Prize. Note that both tasks must be completed. There is no “partial award” for completing just one task. The first person to email me the answers will be awarded the $5,000 prize (greg.carey@colorado.edu). In the event that more than one person emails me satisfactory answers on the same day, the award will be evenly split.

In addition to specifying the tasks, the notes below provide some background, notation, and suggestions on how to approach the problems. I urge you to begin with Task 1. There is no need for any complex algebra to complete this task. You can try finding a solution by simply entering numbers into an Excel spreadsheet.

Before grappling with the matrix algebra involved in Task 2, I strongly suggest that you look over the slides to my presentation as well as the draft paper.

Task 1: Demonstrate that, analogous to common and unique environments, “common genotype” and “unique genotype” can function as variables in a path analysis for full siblings in an intraclass relationship. The proof may be analytic, but to complete this task successfully all that is needed is to find real numbers for variables $A_C$, $A_U_1$, and $A_U_2$ in the following simplified case: a 100% heritable trait due only to a single locus with two alleles in Hardy-Weinberg equilibrium and with equal allele frequencies.

If the environmental variance component within sib-pairs can legitimately be represented as uncorrelated variables $E_1$ and $E_2$ in path analysis and the environmental variance component between sibs can be represented as variable $C$, then the same could be done with the genetic variance components within and between sibships. We should be able to represent the additive genetic variance component within pairs as the uncorrelated variables $A_U_1$ and $A_U_2$ (i.e., the “unique genotype”). Similarly, the additive genetic variance component between pairs may be written as variable $A_C$ (the “common genotype”) that will be uncorrelated with both $A_U_1$ and $A_U_2$.

The model of gene action and marital assortment will determine the ratio of the variance of $A_C$ to $A_U$. Assume random mating. The variance of $A_U$ will equal 0 for identical twins. For fraternal twins and full sibs the variance of $A_U$ will equal the variance of $A_C$. For half-sibs the variance of $A_U$ will equal three times the variance of $A_C$. For first cousins, the variance of $A_U$ will equal seven times the variance of $A_C$. For all relationships, the total genetic variance will be the same and will equal the quantity $A_C + A_U$. The intraclass correlation equals $A_C/(A_C + A_U)$. 

Assume a quantitative trait with 100% heritability due exclusively to the additive gene effects of a single locus with two alleles, A and a. Let \( p \) denote the frequency of \( A \), and \( q = (1 - p) \), the frequency of \( a \). Let the phenotypic values of \( aa, Aa, AA \) be respectively -1, 0, and 1. Under these conditions, the phenotypic mean will equal \((p - q)\), the phenotypic variance will equal \(2pq\), the phenotypic covariance for full siblings will be \(pq\), and the variances of \(A_c, A_{U1}, \) and \(A_{U2}\) will equal \(pq\).

Your task is to find numerical values for \(A_c, A_{U1}, \) and \(A_{U2}\) such that the mean phenotypic value is \((p - q)\), the total phenotypic variance (which, because of 100% heritability, will equal the total genetic variance) equals \(2pq\) and the full sibling covariance matrix for \(A_c, A_{U1}, \) and \(A_{U2}\) takes the form

\[
\begin{pmatrix}
pq & 0 & 0 \\
0 & pq & 0 \\
0 & 0 & pq
\end{pmatrix}
\]  

(1.1)

For the prize, it is sufficient to find these quantities when \(p = q = 0.5\).

The following table will help you get started.

<table>
<thead>
<tr>
<th>Group</th>
<th>Frequency</th>
<th>Genotype:</th>
<th>Phenotypic Value:</th>
<th>(A_c)</th>
<th>(Sib) 1:</th>
<th>(Sib) 2:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(p^2(p + .5q)^2)</td>
<td>AA</td>
<td>AA</td>
<td>1 1</td>
<td>(U)</td>
<td>(1 - U)</td>
</tr>
<tr>
<td>2</td>
<td>(.5p^2q(1 + p))</td>
<td>AA</td>
<td>Aa</td>
<td>1 0</td>
<td>(V)</td>
<td>(1 - V)</td>
</tr>
<tr>
<td>3</td>
<td>(.25p^2q^2)</td>
<td>AA</td>
<td>aa</td>
<td>1 -1</td>
<td>(W)</td>
<td>(1 - W)</td>
</tr>
<tr>
<td>4</td>
<td>(.5p^2q(1 + p))</td>
<td>Aa</td>
<td>AA</td>
<td>0 1</td>
<td>(V)</td>
<td>-V</td>
</tr>
<tr>
<td>5</td>
<td>(pq(1 + pq))</td>
<td>Aa</td>
<td>Aa</td>
<td>0 0</td>
<td>(X)</td>
<td>-X</td>
</tr>
<tr>
<td>6</td>
<td>(.5pq^2(1 + q))</td>
<td>Aa</td>
<td>aa</td>
<td>0 -1</td>
<td>(Y)</td>
<td>-Y</td>
</tr>
<tr>
<td>7</td>
<td>(.25p^2q^2)</td>
<td>aa</td>
<td>AA</td>
<td>-1 1</td>
<td>(W)</td>
<td>-1 - W</td>
</tr>
<tr>
<td>8</td>
<td>(.5pq^2(1 + q))</td>
<td>aa</td>
<td>Aa</td>
<td>-1 0</td>
<td>(Y)</td>
<td>-1 - Y</td>
</tr>
<tr>
<td>9</td>
<td>(q^2(q + .5p)^2)</td>
<td>aa</td>
<td>aa</td>
<td>-1 -1</td>
<td>(Z)</td>
<td>-1 - Z</td>
</tr>
</tbody>
</table>

There are six unknowns in this table, denoted by the letters \(U\) through \(Z\). The assumption of 100% heritability guarantees that the values of the unique genotypic variables for sibs 1 and 2 are a function of these six unknowns. The Excel spreadsheet SingleGene1.xls allows you to enter a value of \(p\) along with those of the six unknowns to see if your guesses provide the correct answers for the phenotypic mean, the phenotypic variance, and the covariance matrix for \(G_c, G_{U1}, \) and \(G_{U2}\).

I do not suggest that you start with this spreadsheet, however, because the problem can be made much simpler by assuming that \(p = .5\). Under this special
condition, it is logical to let $Z = -U$ and $Y = -V$, reducing the number of parameters to four. You may also want to test the situation in which $X$ and $W$ are fixed at 0. This special situation is provided in Excel spreadsheet SingleGene2.xls; the only parameters you need enter are $U$ and $V$.

You can also reduce the problem to one parameter by noting that the mean of $G_C$ will be 0 in this case reducing the equation for the variance of $G_C$ to

$$\text{var}(G_C) = pq = .25 = .28125 * U^2 + .375V^2.$$ 

Just pick a value for $U$ and derive $V$ from this equation.

Finally, if you are truly adventuresome, you will recognize that there is a closed form solution because the variance of $G_{U1}$ and $G_{U2}$ must also equal $pq$. Given that $p$ is .5, the means of $G_{U1}$ and $G_{U2}$ will be 0 and the equation reduces to

$$.25 = .03125 + .28125*(1-U)^2 + .1825(1-2V + 2V^2).$$

I suggest that you brush up on the algebra of complex numbers if you want to pursue this algebraic solution.
Task 2: Demonstrate analytically that the common environmental variable \(C\) and the two unique environmental variables for sibs 1 and sib 2 in an intraclass relationship \(E_1\) and \(E_2\) can be explained as a linear function of concrete environmental events \(Xs\) such that the three variables are uncorrelated. The analytical proof must have the following three properties: (1) it must be nontrivial; (2) it must have reasonable assumptions; and (3) it must lead to a different conclusion than the one given in my 2009 BGA presentation—i.e., the derivation must demonstrate that a randomly chosen \(X\) can have a sibling correlation other than 0 or 1.

Let \(X_{ijk}\) denote the kth concrete environmental variable for the jth member \((j = 1, 2)\) of the ith pair. Then \(X\) can denote the matrix of environmental scores,

\[
X = \begin{pmatrix}
X_{111} & \cdots & X_{11k} & X_{121} & \cdots & X_{12k} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
X_{n11} & \cdots & X_{n1k} & X_{n21} & \cdots & X_{n2k}
\end{pmatrix}
\] (2.1)

Let vector \(y_i\) denote the vector composed of the common environmental variables for the first member of the ith sib pair \((C_i1)\), the unique environmental variable for the first member \((E_i1)\), the common environment for the second member \((C_i2)\), and the unique environment of the second member \((E_i2)\). Then matrix \(Y\) may be written as the horizontal concatenation of the \(y_i\) vectors:

\[
Y = \begin{pmatrix}
y_1 \\
\vdots \\
y_n
\end{pmatrix} = \begin{pmatrix}
C_{i1} & E_{i1} & C_{i2} & E_{i2} \\
\vdots & \vdots & \vdots & \vdots \\
C_{ni} & E_{ni} & C_{n2} & E_{n2}
\end{pmatrix}
\] (2.2)

We can write \(Y\) as a linear function of \(X\) and a matrix of weights, \(B\),

\[
Y = XB
\] (2.3)

where \(B\) may be written as

\[
B = \begin{pmatrix}
b_{C_{111}} & b_{E_{111}} & b_{C_{211}} & b_{E_{211}} \\
\vdots & \vdots & \vdots & \vdots \\
b_{C_{11k}} & b_{E_{11k}} & b_{C_{21k}} & b_{E_{21k}} \\
b_{C_{121}} & b_{E_{121}} & b_{C_{221}} & b_{E_{221}} \\
\vdots & \vdots & \vdots & \vdots \\
b_{C_{12k}} & b_{E_{12k}} & b_{C_{22k}} & b_{E_{22k}}
\end{pmatrix}
\] (2.4)

where \(b_{Cijk}\) denotes the weight given to the kth \(X\) variable of the jth member \((j = 1, 2)\) on the common environmental for the ith member \((i = 1, 2)\).
Assuming that all variables in both $X$ and $Y$ are expressed as deviations from the mean, then the covariance matrix for the variables in $Y$ is

$$\frac{1}{N} Y'Y = \frac{1}{N} B'X'XB$$  \hspace{0.5cm} (2.5)$$

The second task of the price is to provide an analytical form for matrix $B$ such that the covariance matrix for $Y$ takes the following form:

$$\frac{1}{N} Y'Y = \begin{pmatrix}
\sigma_C^2 & 0 & \sigma_C^2 & 0 \\
0 & \sigma_E^2 & 0 & 0 \\
\sigma_C^2 & 0 & \sigma_C^2 & 0 \\
0 & 0 & 0 & \sigma_E^2
\end{pmatrix}$$  \hspace{0.5cm} (2.6)$$

The analytical solution must have three properties. First, it must be nontrivial. A trivial example would let all elements of $B$ equal 0 so that $\sigma_C^2 = \sigma_E^2 = 0$. Similarly, assuming that $\sigma_C^2 = 0$ is also trivial.

The second property is that the form of $B$ must lead to a different conclusion than the one stated in my 2009 BGA presentation—namely, at a single random $X$ must have a sib correlation of either 0.0 or 1.0. You must be able to prove that an $X$ can have a correlation other than 0 or 1 while still preserving the orthogonal nature of variables $C, E_1, \text{and } E_2$.

The final property is that the assumptions must be reasonable. It is perfectly acceptable to make customary simplifying assumptions (e.g., random mating, no gene-environment correlation or interaction, only additive genetic variance). You cannot, however, just “assume the problem away.” For example, it is unreasonable to assume that for every behavioral phenotype, some sibling correlations among the $X$s are positive and some are negative such that they all cancel one another out.

To help you get started, let $X_1$ denote the matrix of all $X$ values for sib 1 and $X_2$, the analogous matrix for sib 2. Then $X$ can be written as the partitioned matrix

$$X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}.$$  \hspace{0.5cm} (2.7)$$

Given an intraclass relationship, the variance-covariance matrix for $X$ can be written as the partitioned matrix

$$\text{cov}(X,X) = \begin{pmatrix} W & S \\
S & W \end{pmatrix}$$  \hspace{0.5cm} (2.8)$$

where $W$ is the covariance matrix of the $X$s within an individual and $S$ is the cross sibling covariance matrix.
Let $b_{Cij}$ denote the column vector of paths from the $X$s of sib $j$ into variable $C$ of the $i$th sib and let $b_{Eij}$ denote the analogous paths from the $X$s of sib $j$ into the variable $E$ of sib $i$. Then matrix $B$ can be written as the partitioned matrix

$$
B = \begin{pmatrix}
b_{C11} & b_{C12} & b_{E11} & b_{E12} \\
b_{C12} & b_{C21} & b_{E12} & b_{E21}
\end{pmatrix}.
$$

Then the desired covariance matrix in Equation 2.6 can be written as

$$
\begin{pmatrix}
b_{C11}' & b_{C12}' \\
b_{E11}' & b_{E12}'
\end{pmatrix}
\begin{pmatrix}
W & S \\
S & W
\end{pmatrix}
\begin{pmatrix}
b_{C11} & b_{E11} & b_{C21} & b_{E21} \\
b_{C12} & b_{E12} & b_{C22} & b_{E22}
\end{pmatrix}
$$

(2.10)

Tedious algebra will reveal that each element in the predicted matrix for Equation 2.6 is a sum of four quadratic forms. (A quadratic form here equals a row vector times a symmetric matrix postmultiplied by a column vector.) Each quadratic form gives a single number. For example, the covariance between $C_1$ and $E_1$ will equal

$$
cov(C_1, E_1) = 0 = b_{C11}'Wb_{E11} + b_{C11}'Sb_{E12} + b_{C12}'Sb_{E11} + b_{C12}'Wb_{E12}
$$

(2.11)

At this stage, you must start making assumptions. It is possible to assume that the four numbers on the right hand side of Equation 2.11 differ from 0 but cancel one another out. This assumption, albeit a logical possibility, is unreasonable. What hidden property in nature would force the elements of the $b$ vectors and the $W$ and $S$ to take on values such that above equation must always be 0 for each and every behavioral phenotype?

A more reasonable situation is to develop the model so that it requires that each of these four numbers equals 0. For example, assume that every $X$ for sib 1 than has an arrow going into $E_1$ is uncorrelated with every $X$ for sib 2 that has an arrow going into $E_2$. This is a reasonable assumption because $E_1$ must be uncorrelated with $E_2$. (The alternative is to make another “magical cancellation” assumption—i.e., that the positive correlations are cancelled out by negative correlations.)

Now order the variables into two mutually exclusive sets. The first, designated as set $T$ includes any $X$ variable that has at least one nonzero correlation between sibs. That is, any $X_i$ for sib 1 that has at least one nonzero correlation with any $X_j$ for sib 2. Then, by definition the second set—denoted here as set $U$—includes all those $X$ variables for sib 1 that correlate exactly 0 with all $X$ variables for sib 2. Because of the intraclass relationship, sib 2’s $X$ variables will have the same numerical indices and those for sib 1. Order the columns in $X_1$ and $X_2$ such that the
$T$ variables are first followed by the $U$ variables. Then matrix $S$ must have the following form:

$$S = \begin{pmatrix} S_T & 0 \\ 0 & 0 \end{pmatrix}$$

Similarly, a vector of weights such as $b_{E12}$ will be partitioned as

$$b_{E12} = \begin{pmatrix} b_{E_T^{12}} \\ b_{E_U^{12}} \end{pmatrix}$$

Then, the quantity $b'_{C1} S b_{E12}$ becomes

$$b'_{C1} S b_{E12} = \begin{pmatrix} b'_{C_T^{11}} & b'_{C_U^{11}} \end{pmatrix} \begin{pmatrix} S_T & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} b_{E_T^{12}} \\ b_{E_U^{12}} \end{pmatrix}$$

so that

$$b'_{C1} S b_{E12} = b'_{C_T^{11}} S_T b_{E_T^{12}} = 0$$

I will leave it to you to find a necessary assumption required for this quantity to be zero. (Hint: if $b_{E_T^{12}} \neq 0$, then what will be the correlation between $E_1$ and $E_2$?). By tediously examining each and every quadratic form that comes from the matrix multiplication involved in Equation 2.10, then ordering the variables into sets $T$ and $U$, you should be able to arrive at the conclusion that I did—any $X$ for sib 1 and its analogous $X$ for sib 2 must correlate either 1 or 0.