# Covariance decomposition in undirected Gaussian graphical models

By
BEATRIX JONES

*Institute of Information and Mathematical Sciences, Massey University, Private Bag 102-904, North Shore Mail Centre, Auckland, New Zealand*
m.b.jones@massey.ac.nz

and MIKE WEST

*Institute of Statistics and Decision Sciences*
*Duke University, Durham, North Carolina 27708-0251, U.S.A.*
mw@stat.duke.edu

Summary

The covariance between two variables in a multivariate Gaussian distribution is decomposed into a sum of path weights for all paths connecting the two variables in an undirected independence graph. These weights are useful in determining which variables are important in mediating correlation between the two path endpoints. The decomposition arises in undirected Gaussian graphical models and does not require or involve any assumptions of causality. This covariance decomposition is derived using basic linear algebra. The decomposition is feasible for very large numbers of variables if the corresponding precision matrix is sparse, a circumstance that arises in examples such as gene expression studies in functional genomics. Additional computational efficiencies are possible when the undirected graph is derived from an acyclic directed graph.

*Some key words*: Concentration graph; Conditional independence; Covariance selection; Path analysis

## 1 Introduction

Graphical models over undirected graphs (Lauritzen, 1996) are increasingly used to exhibit the conditional independence structures in multivariate distributions, and advances in computational techniques are providing increased access to methods associated with graphical modelling in problems of increasing complexity and dimension (Dobra et al., 2004; Rich et al., 2005). Undirected Gaussian graphical models are a special case of particular interest (Speed & Kiiveri, 1986; Giudici, 1996; Jones et al., 2005). Wermuth (1976) showed that, for Gaussian models, conditional independence corresponds to nonzero entries in the precision matrix; thus model selection for undirected Gaussian graphical models is equivalent to selecting which elements of the precision matrix are zero, the problem called 'covariance selection' in Dempster (1972). In exploring and aiming to interpret relationships exhibited in such multivariate systems, we are often faced with questions about how different subsets of possibly many variables mediate the observed relationship between a pair of variables. We represent the covariance between two random variables in a multivariate Gaussian distribution in terms of sums of components related to individual paths between the variables in an underlying graphical model. This decomposition directly defines path weights that highlight and quantify the roles played by intervening variables along multiple such paths.

The covariance decomposition we present is derived using basic linear algebra, relying only on analytical expressions for matrix determinants and inverses. The linear algebra literature has used similar approaches, e. g. to provide alternative expressions for the determinant of a matrix (Johnson et al., 1994); however, to our knowledge they have not been used to aid in the interpretation of covariance patterns in graphical models. We envisage that this decomposition will provide an alternative to the path coefficients of Wright (1921), which provide covariance decomposition along paths between two variables in the context of directed graphs.

## 2 Covariance decomposition over paths

**Theorem 1.** *Consider an $n$-dimensional multivariate distribution with a finite and nonsingular covariance matrix $\Sigma$, with precision matrix $\Omega = \Sigma^{-1}$. Let $\Omega$ determine the incidence matrix of a finite, undirected graph on vertices $(1, \ldots n)$, with nonzero elements in $\Omega$ corresponding to edges. The element of*

*$\Sigma$ corresponding to the covariance between variables $x$ and $y$ can be written as a sum of path weights over all paths in the graph between $x$ and $y$:*

$$\sigma_{xy} = \sum_{P \in \mathcal{P}_{xy}} (-1)^{m+1} \omega_{p_1 p_2} \omega_{p_2 p_3} \dots \omega_{p_{m-1} p_m} \frac{\det(\Omega_{\setminus P})}{\det(\Omega)}, \tag{1}$$

*where $\mathcal{P}_{xy}$ represents the set of paths between $x$ and $y$, so that $p_1 = x$ and $p_m = y$ for all $P \in \mathcal{P}_{xy}$, and $\Omega_{\setminus P}$ is the matrix with rows and columns corresponding to variables in the path $P$ omitted, with the determinant of a 0 dimensional matrix taken to be 1.*

The proof of the theorem uses the following basic results from linear algebra.

*Lemma 1. Let $A$ be a nonsingular, $n$ by $n$ matrix. Then the following hold:*

*(i)*

$$(A^{-1})_{i,j} = \frac{(-1)^{i+j} \det(A_{\setminus j, \setminus i})}{\det(A)}$$

*where the $A_{\setminus j, \setminus i}$ is the matrix produced by removing row $j$ and column $i$ from $A$ (Lang, 1987, Theorem 6.2.3)*

*(ii) for any row $i$:*

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{\setminus i, \setminus j})$$

*with a similar formula based on expanding along a column (Lang, 1987, Theorem 6.8.1).*

As $\Sigma = \Omega^{-1}$, and is symmetric, application of Lemma 1 (i) gives $\sigma_{xy} = (-1)^{x+y} \det(\Omega_{\setminus x, \setminus y}) / \det(\Omega)$. Using Lemma 1 (ii), expand $\det(\Omega_{\setminus x, \setminus y})$ along the column corresponding to the variable $x$. Let $d(\Omega, xi, xy)$ be the appropriate power of $-1$ times the determinant of the matrix, where we have eliminated the rows for variables $x$ and $i$ and the columns for $x$ and $y$. Then

$$\sigma_{xy} = \frac{1}{\det(\Omega)} (\omega_{x1} d(\Omega, x1, xy) + \dots + \omega_{xn} d(\Omega, xn, xy)).$$

There is no term with $\omega_{xx}$ because $\omega_{xx}$ is eliminated from $\Omega$ in producing $\Omega_{\setminus x, \setminus y}$. Terms in the sum drop out for cases with $\omega_{ix} = 0$, i.e. when there is

no edge between variables $i$ and $x$ in the corresponding graph. The nonzero terms can then be expanded further, in terms of edges incident at $i$ in the graph with $x$ eliminated. The first term, representing paths whose first edge is from $x$ to 1, becomes

$$\omega_{x1} d(\Omega, x1, xy) = \omega_{x1}\omega_{12} d(\Omega, x12, xy1) + \ldots + \omega_{x1}\omega_{1n} d(\Omega, x1n, xyn),$$

where $d(\Omega, xij, xyi)$ represents the determinant of the matrix where we have eliminated the rows and columns corresponding to the subscripts, and also subsuming the signs in this term.

Continue the expansion until either an edge to $y$ is reached or the last variable used for the expansion has no remaining incident edge. In the latter case, the determinant of the remaining matrix is zero, and the term does not contribute to $\sigma_{xy}$. The former case constitutes a path from $x$ to $y$; the remaining $d(\cdot)$ term is the determinant of $\Omega$ with the variables in the path removed.

We can determine the relevant power of $-1$ for each path by considering $\Omega$ ordered so that $x$ corresponds to the first row and column, $y$ corresponds to the $m$th row and column, and the intervening rows and columns correspond to the ordered path variables. Under this ordering $(-1)^{x+y} \det(\Omega_{\setminus x,\setminus y})$ has coefficient $(-1)^{m+1}$. The repeated applications of Lemma 2 (ii) to produce the path weight each involve the first row and first column, so the power of $-1$ is not altered. This argument generalises to any ordering of the variables: changing the index of a particular variable involves a row swap and a column swap, and thus does not alter the relevant determinants. □

Let the summand in (1) corresponding to a particular path be called the path weight. An alternative representation gives further insight into what this weight represents. If $X$ is the vector of variables under consideration, the path weight can be written in terms of partial correlations and partial variances of the path variables $X_P$, using Corollaries 5.8.1 and 5.8.2 from Whittaker (1990), and following his notation:

$$\left\{ \prod_{i;p_i \in P\setminus y} \mathrm{cor}(X_{p_i}, X_{p_{i+1}} | X_{\setminus\{p_i,p_{i+1}\}}) \right\} \left\{ \mathrm{var}(X_x | X_{\setminus x}) \mathrm{var}(X_y | X_{\setminus y}) \right\}^{\frac{1}{2}} \times$$

$$\frac{\det(\Sigma_P)}{\prod_{i;p_i \in P} \mathrm{var}(X_{p_i} | X_{\setminus p_i})}.$$

This rearrangement emphasises the fact that the determinants of the potentially large matrices $\Omega$ and $\Omega_{\setminus P}$ need not be computed. We also see that the path weight has the same sign as the product of partial correlations corresponding to the path edges. The term $\left\{\mathrm{var}(X_x|X_{\setminus x})\mathrm{var}(X_y|X_{\setminus y})\right\}^{1/2}$ is common to all paths, and the remaining terms are scale-invariant, so that the path weights are equivariant to scale multiplication of $X_x$ or $X_y$. The ratio $\det(\Sigma_P)/\{\prod_{i;p_i\in P}\mathrm{var}(X_{p_i}|X_{\setminus p_i})\}$ can be further broken down into terms that reflect the general predictability of each path variable, and the strength of association between the path variables:

$$\left\{\frac{\det(\Sigma_P)}{\prod_{i;p_i\in P}\mathrm{var}(X_{p_i}|X_{\setminus p_i})}\right\} = \left\{\prod_{i;p_i\in P}\frac{1}{1-R^2(p_i)}\right\}\left\{\frac{\det(\Sigma_P)}{\prod_{i;p_i\in P}\mathrm{var}(X_{p_i})}\right\},$$

where $R^2(p)$ is the multiple correlation between $X_p$ and $X_{\setminus p}$, and the second term on the right hand side is the information in the marginal distribution of the path variables against their mutual independence.

We also note that, for variables taken to have mean 0 without loss of generality, Theorem 1 can be derived via consideration of $m_{xy}$, the regression coefficient of variable $x$ on $y$ in their bivariate normal distribution: $\sigma_{xy} = E(xy) = m_{xy}\sigma_{yy}$. If $G$ is the matrix of complete conditional regression parameters on the graph, i. e. $G_{ij} = -\omega_{ij}/\omega_{ii}$ with $G_{ii} = 0$, and $U_y$ is the $p$ by $p$ identity matrix with the diagonal element corresponding to $y$ set to zero, $m_{xy}$ is the $(x, y)$ element of $(I - U_yG)^{-1}$. The decomposition along paths can be obtained by applying Lemma 1, in an analogous manner to the proof above, to obtain an expression for the $(x, y)$ element of $(I - U_yG)^{-1}$ in terms of the elements of $G$.

## 3 An illustrative synthetic example

Consider a four-variable example in which $X_{x,a,b,y}$ has precision matrix

$$\Omega = \begin{pmatrix} 5 & 3 & -1.5 & 0 \\ 3 & 5 & -0.5 & 2 \\ -1.5 & -0.5 & 5 & -2.5 \\ 0 & 2 & -2.5 & 5 \end{pmatrix} \tag{2}$$

corresponding to the graph in Fig. 1.

There are three paths from $a$ to $b$, with weights given in column 2 of Table 1; the weights of the four paths from $x$ to $y$ are also given. The weights

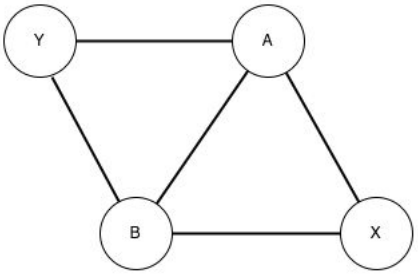


Figure 1: Undirected graph corresponding to the precision matrix in equation (2).

indicate that the direct path between $a$ and $b$ does not contribute as much to the covariance between $a$ and $b$ as do the indirect paths via $x$ and $y$. In fact, the direct path has sign opposite to that of the marginal covariance. We refer to paths whose weights have opposite sign to the marginal covariance as 'moderating' paths, as they reduce or 'moderate' the covariance. Paths with the same sign as the marginal covariance are referred to as 'mediating' paths. The terms 'mediating' and 'moderating' are not used to imply a specific causal relationship: any relationship compatible with the undirected graph is possible, including relationships involving unobserved variables. Nevertheless, the breakdown of the path weights implies that study of the variables $x$ and $y$ is important in understanding the covariance between $a$ and $b$.

Columns 3 and 4 of Table 1 give alternative weights based on decomposition of the correlation and bivariate regression coefficients $m_{xy}$ and $m_{ab}$. These are constant multiples of the covariance based weights, but may be preferable in certain circumstances. Unlike the covariance-based weights, the correlation-based weights are comparable between paths with different endpoints; for example, we can conclude the moderating path $ab$ has larger influence than either moderating path between $x$ and $y$. The weights based on the bivariate regression coefficients indicate how a unit change in $b$ is 'transmitted' to $a$. When $a$ and $b$ are quantities with an intuitive scale, such as survival time in months and weight in pounds, this decomposition may be more easily interpretable.

| | weight based on decomposition of: | | |
|---|---|---|---|
| path | covariance | correlation | reg. coef. |
| $a - b$ | 0.084 | 0.178 | 0.209 |
| $a - x - b$ | −0.151 | −0.320 | −0.375 |
| $a - y - b$ | −0.168 | −0.356 | −0.417 |
| total | −0.235 | −0.500 | −0.584 |
| $x - a - y$ | 0.201 | 0.401 | 0.416 |
| $x - b - y$ | 0.126 | 0.251 | 0.261 |
| $x - a - b - y$ | −0.025 | −0.050 | 0.052 |
| $x - b - a - y$ | −0.010 | −0.020 | 0.021 |
| total | 0.292 | 0.582 | 0.604 |

Table 1: Example. Path weights for all paths between $a$ and $b$, and all paths between $x$ and $y$.

## 4 Undirected graphs derived from acyclic directed graphs

Undirected graphical models are sometimes derived from acyclic directed graphical models. Fitting directed graphical models is often a computationally efficient way of exploring a large subset of the space of undirected graphs. Dobra & West (2004) uses this approach for a data set with over 12,000 variables; marginal likelihood computations for unrestricted, undirected graphs of even 150 variables are infeasible (Jones et al., 2005). Even when computation is not a limiting factor, it may be appealing to consider a fitted acyclic directed graph as an undirected graph. An inferred covariance matrix is often consistent with many acyclic directed graphs with no a priori preference among these, as discussed in Andersson et al. (1997).

Converting an acyclic directed graph to an undirected graph requires adding edges between the parents of each node, called 'moralisation' (Cowell et al., 1999, §3.2), and removing the directionality of the original edges. The endpoint nodes of a moralized edge are dependent conditional on one or more of their common descendants. However, if no common descendants are conditioned upon, the endpoints are independent conditional on some, possibly empty, set of variables. This structure, induced by the underlying acyclic directed graph, implies a specific relationship between certain path weights. We use the following result to find 'cancelling sets' of paths: sets of paths whose weights add to zero.

*Theorem 2. Consider a vector of random variables $X$ with an associated precision matrix and graph. Let $\mathcal{C}$ be the set of paths in this graph from $x$ to $y$ that do not use any variables from the set $A$; these paths also exist after conditioning on the variables in $A$ to produce $X_{\setminus A}|X_A$. Let $D$ be the variables other than $x$ and $y$ used in $\mathcal{C}$. Then the following statements hold.*

*(i) If $(X_A \perp\!\!\!\perp X_D)|X_{xy}$, the path weights for paths in $\mathcal{C}$ sum to zero if and only if they also sum to zero when computed from the precision matrix for $X_{\setminus A}|X_A$.*

*(ii) Define a set of paths $\mathcal{C}'$ from $p_1$ to $p_m$ constructed by using an identical route $p_1, \ldots p_k, x$ for each, then using each path in $\mathcal{C}$ to go from $x$ to $y$, and again identical routes $y, p_{k+1} \ldots p_m$; if $\mathcal{C}$ is a cancelling set and $(X_F \perp\!\!\!\perp X_D)|X_{xy}$, $\mathcal{C}'$ is also a cancelling set.*

The proof of Theorem 2 is given in the Appendix.

Consider applying Theorem 2(i) in the case where $x$ and $y$ are the endpoints of a moralized edge. The two endpoints are independent given some, possibly empty, set of variables, drawn from variables that are ancestors to $x$, $y$, or both. Use this set of variables as $A$ and construct $\mathcal{C}$ and $D$; the one-edge path $xy$ is an element of $\mathcal{C}$. The set $D$ contains the common descendants of $x$ and $y$. After conditioning on $A$, $x$ and $y$ have covariance zero, so $\mathcal{C}$ is a cancelling set for $X_{\setminus A}|X_A$. If, additionally, $X_A$ and $X_D$ are independent conditional on $X_{xy}$, $\mathcal{C}$ will also be a cancelling set for $X$. Theorem 2 (ii) implies that, if a moralised edge $xy$ is a member of a cancelling set, paths with this edge embedded are also members of a cancelling set if the other path vertices $F$ are independent of the vertices in $D$ conditional on variables $x$ and $y$. In the example that follows, we found that the conditions for applying Theorem 2 were frequently satisfied for moralized edges and paths with these edges embedded, greatly reducing the number of paths that needed to be considered.

## 5 An application to analysis of gene expression data

A high-dimensional example arises from exploration of aspects of large-scale graphical models developed for analyses of gene expression data in brain cancer genomics. The data is taken from Rich et al. (2005) and consists of 8408 variables, each measuring the expression level in tumor tissue of a particular gene. The variables are observed in 41 glioblastoma patients, and have undergone univariate transformations so that they have roughly a multivariate Gaussian distribution. The graph is the maximum posterior probability graph found with the method in Dobra & West (2004). We also follow their prior specification and use the maximum a posteriori estimate of $\Omega$ for this graph in computing the weights.

A subgraph of the 8408 variable graph is shown in Fig. 2. The directions from the original acyclic directed graph are included so it is clear which edges are added during the conversion to an undirected graph. Note that for many edges depicted with direction, other graphs in this equivalence class will have that edge pointing in the opposite direction. Suppose our interest is in whether the expression of gene KIAA0913, which encodes a protein of unknown function, is worth investigating as a potential intermediary between the genes ZBP1 and TGM1. The expression levels for ZBP1 and TGM1 have covariance 0.68, correlation 0.78. All paths considered are between the same two endpoints, so we use the covariance based weights.

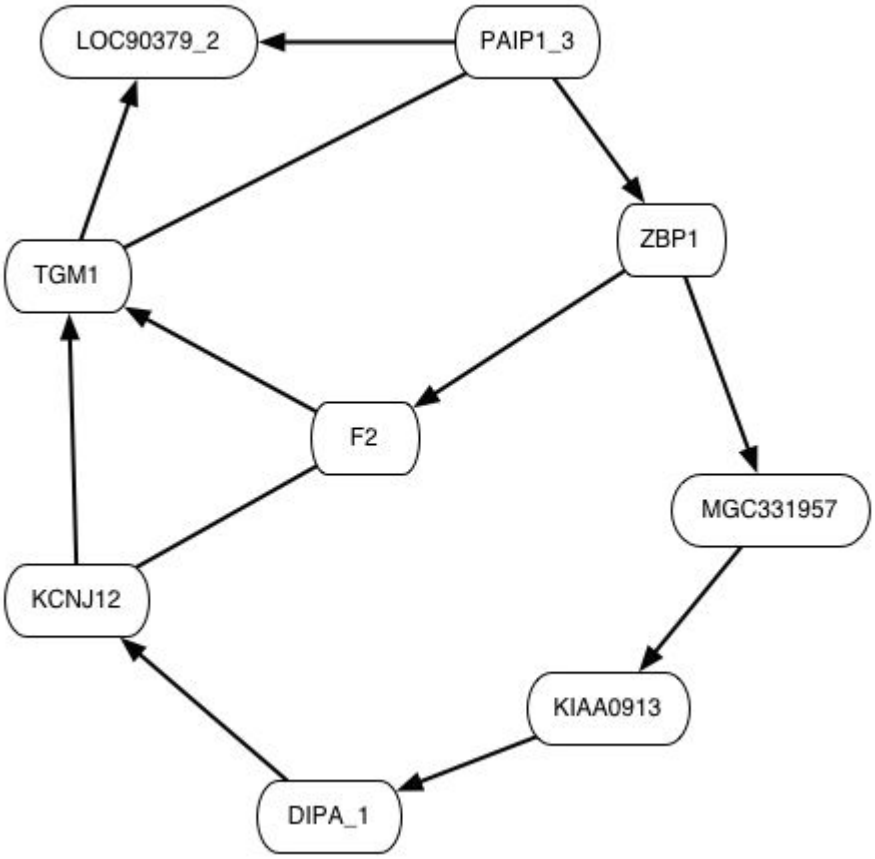


Figure 2: Subgraph of an 8408-node graph representing gene expression in glioblastomas.

The paths ZBP1-PAIP1_3-TGM1 and ZBP1-PAIP1_3-LOC90379_2-TGM1 have path weights of the exactly the same magnitude and opposite sign. This is a consequence of Theorem 2. If the acyclic directed graph actually represents the causal structure, in otherwords, the endpoints $x$ and $y$ of the moralised edge are independent influences on their common offspring, the moralised edge and paths via the common descendants of $x$ and $y$ are not involved in transmitting influence between $x$ and $y$. Consequently we have omitted cancelling sets of paths from consideration. This dramatically reduces the number of paths to be considered; in this example more than 60 paths, not shown, were eliminated.

| Path via | Weight |
|---|---|
| F2 | 0.97 |
| F2-KCNJ12 | −0.56 |
| MGC31957-KIAA0913-DIPA_1-KCNJ12 | 0.43 |
| MGC31957-KIAA0913-DIPA_1-KCNJ12-F2 | −0.16 |
| Total | 0.68 |

Table 2: Gene expression application. Paths contributing to the covariance between ZBP1 and TGM1. All paths start at ZBP1 and end at TGM1, so these endpoints have been omitted.

The four paths that are not part of cancelling sets have weights given in Table 2. Note that not all paths involving moralised edges can be disregarded; for example, the path via F2 and KCNJ12 is an important moderating path. We also see that the paths through KIAA0913 do make an important contribution, accounting for about 40% of the covariance between ZBP1 and TGM1. Finally, despite the $\Omega$ matrix being of dimension 8408 x 8408, because this matrix is sparse, and never needs to be dealt with as a whole, the calculation of these path weights takes less than a minute in Matlab.

## Acknowledgement

Partial support for the research described here was provided by the Keck Foundation through the Keck Center for Neurooncogenomics at Duke University, and the U. S. National Science Foundation. We would also like to acknowledge Adrian Dobra and Quanli Wang for software for graphical model fitting and visualisation, and the helpful comments of the editor and two referees.

## Appendix

### *Proof of Theorem 2*

Part (i). We show the ratio between the path weight computed for $X_{\setminus A}|X_A$ and a path weight computed for $X$ is constant over paths in $\mathcal{C}$. This implies that the path weights sum to zero over $\mathcal{C}$ either in both cases, or in neither. Since the precision matrix for $X_{\setminus A}|X_A$ is produced from the precision matrix for $X$ by simply eliminating the rows and columns for the variables in $A$, the ratio of weights for a path $P$ containing $m$ vertices is:

$$\frac{(-1)^{m+1}\omega_{p_1,p_2}\dots\omega_{p_{m-1},p_m}\det(\Omega_{\setminus\{P\cup A\}})/\det(\Omega_{\setminus A})}{(-1)^{m+1}\omega_{p_1,p_2}\dots\omega_{p_{m-1},p_m}\det(\Omega_{\setminus P})/\det(\Omega)}$$

Only $\det(\Omega_{\setminus\{P\cup A\}})/\det(\Omega_{\setminus P})$ is a function of $P$; since $\Omega_{\setminus\{P\cup A\}}$ and $\Omega_{\setminus P}$ are positive definite precision matrices, this ratio of determinants can also be written in terms of the matrix inverses, which are conditional covariance matrices:

$$\frac{\det(\Omega_{\setminus\{P\cup A\}})}{\det(\Omega_{\setminus P})}=\frac{\det(\Sigma_{\setminus P|P})}{\det(\Sigma_{\setminus\{P\cup A\}|P\cup A})}=\frac{\det(\Sigma_{A|P})\det(\Sigma_{\setminus\{P\cup A\}|P\cup A})}{\det(\Sigma_{\setminus\{P\cup A\}|P\cup A})}=\det(\Sigma_{A|P}).$$

Since every path contains $x$ and $y$, and $(X_A \perp\!\!\!\perp X_D)|X_{xy}$, $\Sigma_{A|P}$ is constant over $P\in\mathcal{C}$.

Part (ii). From part (i) the path weights for paths in $\mathcal{C}$ will also sum to zero for $X_{\setminus F}|X_F$. The ratio between the path weights for $\mathcal{C}'$ for $X$ and the weights for corresponding paths in $\mathcal{C}$ for $X_{\setminus F}|X_F$ is

$$(-1)^m\omega_{p_1,p_2}\dots\omega_{p_k,x}\omega_{y,p_{k+1}}\dots\omega_{p_{m-1},p_m}\det(\Omega_{\setminus F})/\det(\Omega),$$

which is constant over paths in $\mathcal{C}'$. Thus the weights for paths in $\mathcal{C}'$ computed for $X$ also sum to zero. □