

Stumbling across symmetries

Anna Gottard

Abstract In this work, the class of graphical log-linear models for homologous factors introduced by Gottard (2009) is analyzed. This class of models, called SQS models, allows for different kinds of symmetry constraint on main effects and interactions parameters of a graphical log-linear model. The conditional independence structure of a SQS model is represented by a coloured graph in which constraints on main effects are associated with coloured nodes and constraints on interaction parameters are associated with coloured edges. Quasi-symmetry and complete symmetry models are special cases of SQS models. The aim of this work is to analyze which kind of SQS models are compatible with contingency tables in which some conditional margins are symmetric. The three dimensional case is considered.

Key words: Graphical models, symmetry models, coloured graphs, categorical data analysis.

1 Introduction

Graphical log-linear models, described by Darroch *et al.* (1975), are a subclass of log-linear models, such that each model is associated with an undirected graph. In the graph each factor is represented by a node and the absence of an edge connecting nodes represents conditional independence.

In some applications, it is possible to encounter homologous factors, being measured with the same categorical scale. They are commonly encountered in matched pairs studies, in attitudinal research where subjects opinions about some issue are observed under several conditions, or in longitudinal studies that repeatedly observe whether a condition is present at various times. A variety of contexts in which homologous factors can arise is described in McCullagh (1982). In all these cases, it can happen that the contingency table cross classifying the homologous factors

Anna Gottard
Dipartimento di Statistica dell'Università di Firenze, e-mail: gottard@ds.unifi.it

shows an exchangeable structure. This structure can be typically captured by symmetry and quasi-symmetry models (see Caussinus, 1966, Bishop *et al.*, 1975 and Agresti, 2002). In the first class of models, a symmetric contingency table is assumed, while the latter one supposes a sort of symmetry among the interaction terms of a log-linear model. A larger class of models for exchangeable structures has been introduced by Lovison (2000), in the case of hypercubic concordance tables.

Restricting to graphical models, at the presence of homologous factors Gottard *et al.* (2010) introduced the class of Quasi-Symmetric (QS) graphical models. QS graphical models extend quasi-symmetry models to the case of a generic conditional independence structure and to include non-homologous factors. Moreover, Gottard (2009) defines *Symmetric and quasi-symmetric (SQS) graphical models* to include both quasi-symmetry models and complete symmetry models as a special case. These two classes of model utilize a particular kind of coloured graph to depict both conditional independence restrictions and equality and symmetry constraints among the log-linear parameters.

In this paper, SQS graphical models will be better analyzed to deduce which kinds of graph are compatible with symmetric contingency tables or contingency tables with some symmetric margins.

The paper is organized as follows. Section 2 defines SQS graphical models. We will limit to the case of a three dimensional contingency table. In Section 3 we describe characteristics of the coloured graphs that are compatible with symmetric contingency tables or margins. The last section makes some concluding remarks.

2 Symmetric and quasi-symmetric graphical models

Suppose a three-dimensional discrete random variable $\mathbf{X}_V = X_A, X_B, X_C$, whose components take values in $[I] = \{1, \dots, I\}$, all components being homologous. Let $\mathcal{J} = \prod_{v \in V} [r_v]$ be the contingency table cross-classifying these factors and $\iota = (i, j, k)$ its generic cell. The joint distribution of $\mathbf{X}_V$, under a multinomial sampling of size n , can be specified by a set of probabilities $p(\iota) = P(X_A = i, X_B = j, X_C = k)$, here assumed strictly positive. Moreover, call $n(\iota)$ the observed count in cell ι , and $\mu(\iota) = np(\iota)$ its expected value. The saturated log-linear model for $\mathbf{X}_V$ is then

$$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}. \quad (1)$$

This model assumes all the three factors to be mutually dependent. Deleting some $\lambda^a(\iota_a)$, models admitting conditional or marginal independence can be defined.

Typically, the homologous factors are associated with two classes of model: the symmetry and the quasi-symmetry model. The definition of the first model chronologically precedes the definition of the second, but the former is in fact a reduced, nested model of the latter.

The *quasi-symmetry model* for these three factors can be obtained by imposing the following constraints on interaction parameters of model (1):

$$\lambda_{ij}^{AB} = \lambda_{ji}^{AB} = \lambda_{ij}^{AC} = \lambda_{ji}^{AC} = \lambda_{ij}^{BC} = \lambda_{ji}^{BC} \quad \text{and} \quad \lambda_{ijk}^{ABC} = \lambda_{\text{perm}(ijk)}^{ABC} \quad (2)$$

for each $i, j, k = 1, \dots, I$, where $\text{perm}(ijk)$ denotes all the possible permutations of the indexes.

The *symmetry model* can be obtained adding further constraints to (2), on main effects

$$\lambda_i^A = \lambda_i^B = \lambda_i^C \quad \forall i = 1, \dots, I. \quad (3)$$

As aforementioned, under a symmetry model $\mu_{ijk} = \mu_{\text{perm}(ijk)}$.

Graphical log-linear models are a class of log-linear models in which some interaction terms are deleted according to an undirected graph. An undirected graph $G = (V, E)$ is defined by a set of nodes, V , and a set of edges $E \subset V \times V$. Each node represents a random variable and the absence of an edge $vw = (v, w)$ in E implies conditional independence between X_v and X_w , given the other variables. For instance, whenever the edge between X_A and X_B is absent in the graph, then

$$X_A \perp\!\!\!\perp X_B \mid X_C \quad \Leftrightarrow \quad \lambda_{ij}^{AB} = \lambda_{ijk}^{ABC} = 0 \quad \forall i, j, k.$$

See Lauritzen (1996) for further details.

Consider now the case of X_A, X_B, X_C being homologous. An SQS model has been defined in Gottard (2009) as an extension of the QS graphical models proposed by Gottard *et al.* (2010). QS and SQS models are a particular kind of graphical log-linear models in which it is possible to impose both zero-constraints on parameters to represent conditional independence and equality constraints to take into account some kind of exchangeable structure among homologous factors. Equality constraints are portrayed by adopting a coloured version of an undirected graph. In the following, the definitions of coloured graphs and SQS models are recalled.

Definition 1. A *coloured graph* is a quartet $G = (V, E, \mathcal{V}, \mathcal{E})$ in which:

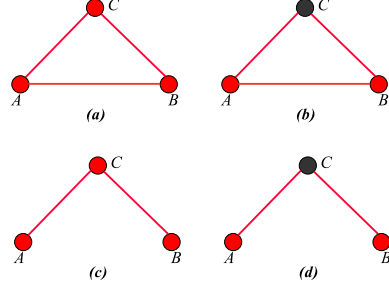
1. V is a set of nodes;
2. E is a set of undirected edges connecting nodes, vw , with both v and w in V ;
3. $\mathcal{V}$ is a partition of V into disjoint sets, called *node-colour classes*, $V_0, \dots, V_r$. Each node-colour class collects nodes with the same colour. The colour black is associated V_0 .
4. $\mathcal{E}$ is a partition of E into disjoint sets, called *edge-colour classes*, $E_0, \dots, E_s$. Each edge-colour class collects edges with the same colour. The colour black is associated with E_0 .

Notice that this definition differs to that given by Hjsgaard and Lauritzen (2008), because of the peculiarity of the discrete case. In particular, all black edges and nodes are not represented as singletons, but allocated altogether into a sole set with index 0.

The black version of a coloured graph G , obtained by replacing all colours by black is called *skeleton graph* and denoted G^* . Hence, the colour black is treated separately from the other colours. The word *colour* will be here used as *not black*.

Example. Consider for example the graph (d) in Figure 1. Here the set of nodes is $V = \{A, B, C\}$, while the set of edges is $E = \{AB, BC\}$. This two sets, ignoring

Fig. 1 Graphs compatible with symmetry: (a) complete symmetry table; (b), (c) and (d) contingency table with symmetric AB margins and non symmetric AC and BC margins.



colours, define the skeleton graph. The set of node-colour classes is $\mathcal{V} = \{V_0, V_1\}$, in which $V_0 = \{C\}$ and $V_1 = \{A, B\}$. As all the edges are coloured, $\mathcal{E} = \{E_1\}$.

We can introduce now the definition of SQS graphical models as follows.

Definition 2. A SQS graphical model is a log-linear model with associated coloured graph $G = (V, E, \mathcal{V}, \mathcal{E})$, with parameters $\lambda^a(i_a)$ for $a \subseteq V$, such that:

1. the model satisfies the conditional independence constraints of the graphical model associated with the skeleton graph G^* ;
2. a coloured edge denotes that the two-factor parameters are set equal with respect to index permutation;

$$\lambda^{vw}(ij) = \lambda^{vw}(ji), \quad \forall i, j = 1, \dots, I$$

with $v, w \in V$. Moreover the nonzero three-factor parameters λ_i^V satisfy the quasi-symmetry equality constraints of the type

$$\lambda_{ijk}^V = \lambda_{jik}^V, \quad \forall i, j, k = 1, \dots, I$$

or similar for different index k positions;

3. for all edges C_1 and C_2 of the same colour, the interaction parameters of second order satisfy the equality constraints

$$\lambda_{ij}^{C_1} = \lambda_{ij}^{C_2}, \quad \text{for all } i, j = 1, \dots, I;$$

4. every two nodes a, b in a same node-colour class $V_l, l \neq 0$, the following set of constraints are imposed on the main-effect parameters:

$$\lambda_i^a = \lambda_i^b \quad \text{for all } i = 1, \dots, I$$

Example. As an example for SQS models definition, we can look again at the graph (d) in Figure 1. To define a possible model associated with this graph, let us start by the model in equation (1). As the skeleton graph obtained by replacing the colour red with black has the edge AB missing, for Definition 2.1, the associated SQS model has $X_A \perp\!\!\!\perp X_B \mid X_C$. Hence,

$$\lambda_{ijk}^{ABC} = 0 \quad \text{and} \quad \lambda_{ij}^{AB} = 0 \quad \forall i, j, k = 1, \dots, I.$$

Both the edges are coloured, thus, for Definition 2.2, it is

$$\lambda_{ik}^{AC} = \lambda_{ki}^{AC} \quad \text{and} \quad \lambda_{jk}^{BC} = \lambda_{kj}^{BC} \quad \forall i, j, k = 1, \dots, I$$

and, as they share the same colour, for Definition 2.3, it is

$$\lambda_{ij}^{AC} = \lambda_{ij}^{BC} \quad \forall i, j = 1, \dots, I.$$

Eventually, as the graph considered has two red nodes, namely A and B , and one black node, C , further constraints have to be added on the main effects according to Definition 2.4

$$\lambda_i^A = \lambda_i^B \quad \forall i = 1, \dots, I.$$

This model is neither an ordinary symmetry model nor an ordinary quasi-symmetry model.

3 Symmetric tables and margins

A joint distribution for three homologous factors is said to entail a symmetric contingency table if the expected cell counts satisfy the following equality

$$\mu_{ijk} = \mu_{\text{perm}(ijk)} \quad \forall i, j, k = 1, \dots, I. \quad (4)$$

This condition is satisfied, in a meaningful way for homologous factors, by the ordinary symmetry models.

Similarly, a joint distribution for three homologous factors is said to entail a symmetric margin, say AB , if the expected cell count satisfies

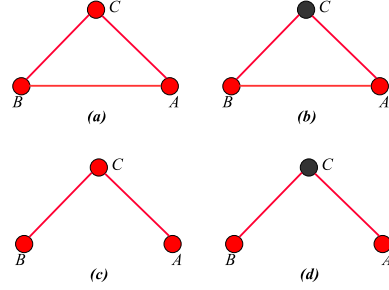
$$\mu_{ijk} = \mu_{jik} \quad \forall i, j, k = 1, \dots, I. \quad (5)$$

Note that this equality implies that all the conditional distributions for $(X_A, X_B) \mid X_C$ and the marginal distribution for (X_A, X_B) entail symmetric contingency tables.

Let us analyze in the first place which kinds of SQS model, and therefore which kinds of coloured graph, conform with a symmetric table. As it is known that the ordinary symmetry model involves a symmetric table, condition (4) is satisfied, whenever the model fulfills the constraints in (2) and (3).

According to Definition 2, constraints (2) are true for all the SQS models such that all the edges share a same colour, i.e. $E_0 = \emptyset$ and $E_1 = E$. Moreover, constraints in (3) are true, whenever all the nodes are coloured by a same colour, that is $V_0 = \emptyset$ and $V_1 = V$. Notice that the edges have to be either all present or all missing. In fact, the symmetry model admits only the complete dependence and the complete independence case.

Fig. 2 Graphs obtained by graphs in figure 1 by switching label A and B.



As a general rule, we can say that an SQS model is compatible with a symmetric contingency table, namely it is *fully symmetric*, whenever both V_0 and E_0 are empty and permuting the labels of the nodes in V , nothing changes in E , $\mathcal{V}$ and $\mathcal{E}$.

In fact, whenever V_0 and E_0 are empty, all the nodes and the edges are coloured. Furthermore, if by permuting the labels of the nodes E , $\mathcal{V}$ and $\mathcal{E}$ remain unchanged, this implies that all the nodes are in a same node-colour class, and all the edges are in the edge-colour class or are missing. Thus, constraints (3) and (2) are satisfied and condition (4) is fulfilled.

In Figure 1, the only graph satisfying this condition is the graph (a). Figures 2 and 3 (a) report the graph obtained by that in Figure 1 by exchanging label A with label B and label A with label C respectively. It can be noted that nothing is changed in the graph. A further graph satisfying condition (4) is a graph with all the edges missing and all the nodes of the same colour, i.e. a fully coloured null graph.

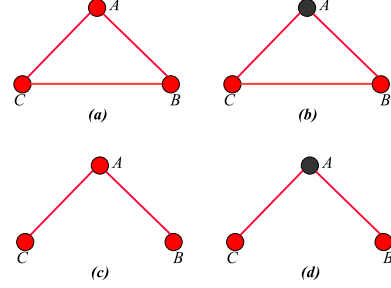
The possibility of conditional, and therefore marginal, tables to be symmetric requires some more argumentations. Consider for instance the margin AB . First of all, notice that $\mu_{ijk} = \mu_{jik}$ implies $\log \mu_{ijk} = \log \mu_{jik}$. This is true whenever

$$\begin{aligned} \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC} \\ = \\ \lambda + \lambda_j^A + \lambda_i^B + \lambda_k^C + \lambda_{ji}^{AB} + \lambda_{jk}^{AC} + \lambda_{ik}^{BC} + \lambda_{jik}^{ABC} \end{aligned}$$

for all $i, j, k = 1, \dots, I$. This equality can be achieved in different ways. We are here considering a solution compatible with a sensible interpretation for homologous factors.

First of all, one can establish that $\lambda_i^A = \lambda_i^B$, $\forall i = 1, \dots, I$ that is, according with Definition 2.4, the two nodes A and B have to share the same colour. Then, one can impose the set constraints $\lambda_{ij}^{AB} = \lambda_{ji}^{AB}$ and for all i, j, k in $1, \dots, I$. In SQS models, this set of constraints is satisfied whenever the edge AB is not black, i.e. either it is coloured or missing. For the second part of Definition 2.2, this coloured edge

Fig. 3 Graphs obtained by graphs in figure 1 by switching label A and C.



implies also that $\lambda_{ijk}^{ABC} = \lambda_{jik}^{ABC}$. Finally, one can impose $\lambda_{jk}^{BC} = \lambda_{jk}^{AC} \forall j, k = 1, \dots, I$, i.e. that the edges AC and BC have a same colour (Definition 2.3). Notice that some further constraints have been actually placed, more than what strictly requested. For example, $\lambda_{ijk}^{ABC} = \lambda_{\text{perm}(ijk)}^{ABC}$. See Proposition 1 in Gottard *et al.* (2010) as an explanation of this aspect.

As a general rule, we can say that an SQS model is compatible with a symmetric margin, namely it is *conditionally symmetric*, whenever the subgraph associated with the considered margin has no nodes in V_0 and no edges in E_0 , and permuting the labels of nodes within the subgraph, nothing changes in E , $\mathcal{V}$ and $\mathcal{E}$.

Substantially, this condition ensures that the subgraph of interest is fully symmetric and that all the edges touching a node in the subgraph have a same colour. In the example above, this second part ensures that $\lambda_{jk}^{BC} = \lambda_{jk}^{AC} \forall j, k = 1, \dots, I$.

Example. Let us take into account graphs in Figure 1. The three graphs (b), (c) and (d) admit the margin AB to be conditionally symmetric. In fact, the subgraph G_{AB} has the two nodes of the same colour, and the edge AB is either coloured or missing. Therefore, the subgraph is fully symmetric. Moreover, permuting label A and label B , the graphs remain unchanged, compare Figure 1 with Figure 2. Consider, instead, the margin AC for the same group of graphs. Both the graphs (b) and (d) have subgraphs G_{AC} not fully symmetric, as the node C is not coloured. On the contrary, the subgraph G_{AC} is fully symmetric in (c). However, permuting the labels A and C (see Figure 3 (c)), the edge BC , present in Figure 1 (c), is missing in Figure 3 (c). Vice versa for the edge AB .

4 Discussion

In this work it has been shown how the use of a coloured graph can be helpful to construct models for contingency tables entailing symmetric frequencies in a contingency table or in its conditional margins. Only the three dimensional case has

been explored, but the extension to a higher dimensionality would comport only an increased complexity on the notation.

Regarding the graphical conditions found to detect symmetries, it seems interesting to notice that the exchangeability assumed among the homologous factors that implies symmetry in the contingency table or in the conditional margins, parallels with a sort of exchangeability among the labels in the graph.

The class of SQS models explored here admits symmetry in the entire contingency table or in all conditional margins, given a certain variable. An interesting expansion concerns the class of context specific interaction models (Højsgaard, 2004) in which symmetries can be involved only for specific levels of a conditional factor.

Acknowledgements This research was partially supported by PRIN 2007, n. 2007XECZ7L002, from MIUR, Italy. The author is grateful to Giovanni Marchetti and Alan Agresti for interesting discussions.

References

1. Agresti, A.: *Categorical Data Analysis*. Second edition, John Wiley & Sons (2002)
2. Bishop, Y.M.M., Fienberg, S.E., Holland, P.W.: *Discrete Multivariate Analysis: Theory and Practice*. MIT Press (1975)
3. Caussinus, H.: Contribution à l'analyse statistique des tableaux de corrélation. *Annales de la Faculté des Sciences de l'Université de Toulouse*, **29**, 77–182 (1966)
4. Darroch, J.N., Lauritzen, S.L., Speed, T.P.: Markov Fields and Log-linear Interaction Models for Contingency Tables. *Ann. Statist.* **8**, 522–539 (1980)
5. Gottard, A.: Graphical log-linear models for homologous factors. In Paganoni, A.M., Sangalli, L.M., Secchi, P., Vantini, S. (eds) *S.Co.2009. Complex data modeling and computationally intensive statistical methods for estimation and prediction*. Proceedings. Politecnico di Milano, September 14 - 16, 2009, pp. 203-208. Maggioli editore. (2009)
6. Gottard, A., Marchetti, G.M., Agresti, A.: Quasi-symmetric graphical log-linear models. (2010) Submitted
7. Højsgaard, S.: Statistical Inference in Context Specific Interaction Models for Contingency Tables. *Scand. J. of Stat.* **31**, 143-158 (2004)
8. Højsgaard, S., Lauritzen, S.L.: Graphical Gaussian models with edge and vertex symmetries. *J. R. Statist. Soc. Ser. B* **70**, 1005-1027 (2008)
9. Lauritzen, S.L.: *Graphical Models*. Oxford: Oxford Science Publications (1996)
10. Lovison, G.: Generalized symmetry models for hypercubic concordance tables. *Int. Stat. Rev.* **68**, 323–338 (2000)
11. McCullagh, P.: Some Applications of Quasisymmetry. *Biometrika* **69**, 303–308 (1982)