Caputo:

Factorization of the Cumulative Distribution Function in Case of Conditional Independence. (REVISED, November 1999)

Sonderforschungsbereich 386, Paper 161 (1999)

Factorization of the Cumulative Distribution Function in Case of Conditional Independence

By Angelika Caputo
University of Munich, Institute of Statistics,
Ludwigstr. 33, D-80539 Munich, Germany

Abstract

A decomposition of complex estimation problems is often obtained by using factorization formulae for the underlying likelihood or density function. This is, for instance, the case in so–called decomposable graphical models where under the restrictions of conditional independences induced by the graph the estimation in the original model may be decomposed into estimation problems corresponding to subgraphs. Such a decomposition is based on the property of conditional independence and on the factorization of the assumed underlying density function. In this paper, analogous factorization formulae for the cdf are introduced which can be useful in situations where the density is not tractable.

Key words: Conditional independence, graphical models, decomposition of ML estimation.

1 Introduction and notations

Highdimensional estimation problems can often be simplified using a factorization of the underlying likelihood function which is based on a corresponding factorization of the density function. The simplification can be obtained if the resulting factors are distinct with respect to the underlying parameter vector. In special cases these factors can be assigned to submodels of the original model. This is, for instance, of concern in the framework of graphical models where a graphical representation is used showing conditional independencies between subvectors (cf. Lauritzen and Wermuth, 1989, Wermuth and Lauritzen, 1990, Lauritzen, 1996). Typically, the underlying multivariate distribution is assumed to be the multivariate normal or the so–called Conditional–Gaussian distribution. Both are members of the exponential family which implies that the multiplicative structure of the density can be exploited. Therefore, they are advantageous with regard to the desired factorization. In addition, a special structure of the graph due to conditional independence

possibly results in a decomposition of the maximum likelihood (ML) estimation of the parameters involved. The possibility of using the graphical representation of the model in the described way is one important benefit of graphical models. If other multivariate distributions are considered, especially if the density cannot be treated analytically, the decomposition is lost either because the density cannot be factorized or because it is impossible to show that the factorization holds although the needed structure of the graph is given. In some cases, however, the cumulative distribution function (cdf) is easy to handle although the density function is not – but up to now no factorization formulae for cdf's are known to the best knowledge of the author.
This paper focusses on a factorization for the cdf which is well–known for the density function itself. The factorization formula for the density function as well as the one for the cdf hold in the case of conditional independence which is typically considered in graphical models.
Because of its prominent role first the concept of conditional independence is introduced followed by some basic notations and definitions from the theory of graphical models. In that what follows, two factorization formulae for the density function are recalled. In Section 2, the main results are presented with focus on dual factorization theorems for the cdf. Section 3 shows the application of these formulae to the estimation problem in graphical models where a multivariate distribution is assumed with intractable density but tractable cdf. In comparison to the more natural way via density or likelihood functions, it is described in detail how the general concept of a cut (Barndorff–Nielsen, 1978) can be used in a modified way for cdf's. The final section deals with a concrete family of multivariate distributions where in fact the density function in contrast to the cdf is not feasible.

Let $X = (X_i)_{i \in V} = (X_1, \dots, X_p)^T$ denote a vector of random variables with related index set $V = \{1, \dots, p\}$ and P_X or briefly P the distribution of X, i.e. the joint distribution of $X_1, \dots, X_p$. The corresponding cdf is given by $F_V(x) = F(x) = F(x_1, \dots, x_p) = P(X_1 \leq x_1, \dots, X_p \leq x_p)$. Let us further assume that the density exists and reads as $f_V(x) = f(x) = f(x_1, \dots, x_p)$. For a subset A of V and the corresponding subvector $X_A = (X_j)_{j \in A}$ let $f_A(x_A)$ and $F_A(x_A)$ denote the density and the cdf of the marginal distribution of X_A, respectively. For disjoint subsets A and B of V the conditional density function of X_A given $X_B = x_B$ is defined as

$$f_{A|B}(x_A|x_B) = \begin{cases} \frac{f_{A \cup B}(x_A, x_B)}{f_B(x_B)}, & f_B(x_B) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

and the cdf of the conditional distribution of X_A given $X_B = x_B$ can be calculated as

$$F_{A|B}(x_A|x_B) = \int_{-\infty}^{x_A} f_{A|B}(u|x_B)du. \tag{1.2}$$

As mentioned above, the concept of conditional independence plays an important role for the persued factorization. In the following, this property is defined similarly to Dawid (1979) by means of a factorization formula for the joint density and for the joint cdf. A more theoretical definition via versions of conditional expectations (Dawid, 1980) is possible but not necessary in this context. Here, the definition close to density functions and cdf's makes the connection to the factorization of these functions obvious. For disjoint subsets A, B, and C of V X_A and X_B are said to be

(i) conditionally independent given $X_C = x_C$ if the conditional density or the cdf of $(X_A^T, X_B^T)^T$ given $X_C = x_C$ is a product of the marginal conditional densities or cdf's. To be more precise, X_A and X_B are conditionally independent given $X_C = x_C$ if it holds for all $x_A \in I\!R^{|A|}$ and all $x_B \in I\!R^{|B|}$

$$
\begin{aligned}
& f_{A\cup B|C}(x_A, x_B|x_C) &=& f_{A|C}(x_A|x_C)f_{B|C}(x_B|x_C) & (1.3)\\
\text{or} \quad & F_{A\cup B|C}(x_A, x_B|x_C) &=& F_{A|C}(x_A|x_C)F_{B|C}(x_B|x_C), & (1.4)
\end{aligned}
$$

(ii) conditionally independent given X_C if X_A and X_B are conditionally independent given $X_C = x_C$ for all $x_C \in I\!R^{|C|}$. Then we write $X_A \perp X_B \mid X_C$ or briefly $A \perp B \mid C$.

The notation $A \perp B \mid C$ was introduced by Dawid (1979, 1980). He discusses basic properties, applications, and possible interpretations of the concept of conditional independence in detail. His ideas are not only of theoretical interest. They are, for instance, recovered in the theory of graphical models. For convenience, let us recall the basic ideas of this important class of multivariate models which will be used later to illustrate the practical relevance of the results derived in this paper. A graph $\mathcal{G} = (V, E)$ is given by a set of vertices V representing the variables and a set of edges $E \subset V \times V$ with $(i,i) \notin E$ for all $i \in V$ reflecting associations among the variables. The set of vertices is identified with the index set of the vector X_V. Here, only so–called undirected graphs where $(i,j) \in E$ implies $(j,i) \in E$ are considered. For $A \subset V$ a subgraph $\mathcal{G}_A$ is defined as $\mathcal{G}_A = (A, E \cap (A \times A))$. A path from i to j is given by a sequence $i = i_0, i_1, \ldots, i_n = j$ with $(i_m, i_{m+1}) \in E$ for $m = 0, \ldots, n-1$. For disjoint subsets A, B, C of V we say C separates A and B if each path from a vertex $i \in A$ to $j \in B$ includes at least one vertex $k \in C$.

In the theory of graphical models, so–called Markov properties connect the concept of conditional independence with graph theory. The distribution of X is called Markov with respect to $\mathcal{G} = (V, E)$ if $A \perp B \mid C$ holds for all disjoint subsets A, B, C of V whenever A and B are separated by the set C in the corresponding graph.
For a partition A, B, and C of V with $A \perp B \mid C$ the general factorization of the density

$$f(x) = f_{A \cup B \cup C}(x_A, x_B, x_C) = f_{A \cup C}(x_A, x_C) f_{B|A \cup C}(x_B | x_A, x_C) \tag{1.5}$$

may be written as

$$f(x) = f_{A \cup C}(x_A, x_C) f_{B|C}(x_B | x_C) = \frac{f_{A \cup C}(x_A, x_C) f_{B \cup C}(x_B, x_C)}{f_C(x_C)} \tag{1.6}$$

for all $x \in I\!R^p$ with $f_C(x_C) \neq 0$. In the theory of graphical models, this formula is basically needed to show a possible decomposition of the ML estimation provided the assumption $A \perp B \mid C$ is fulfilled. In the next section, it is shown that formulae dual to Equations (1.5) and (1.6) hold for the cdf in the special case of conditional independence. This result will then be used for the decomposition of the ML estimation in graphical models in Section 3.

2 Factorization properties

In this section, fundamental properties of the cdf are discussed some of which are direct consequences from the definition of conditional independence or well–known equations like Bayes formula and some look at first sight trivially but only seem to be. First, two formulae for the cdf are derived in the case of the conditional independence $A \perp B \mid C$ for a partition A, B, and C of V. Then, a general property of multivariate distributions is presented which can be used to conclude from the factorization of the cdf or certain conditional propabilities to the conditional independence of A and B given C, namely the inversion of the first formulated statements.
Let $X = X_V$ be a vector of random variables with cdf $F(x)$. We consider a partition A, B, and C of V. Let $\sigma(X_A)$, $\sigma(X_B)$, and $\sigma(X_C)$ be the σ–algebras generated by X_A, X_B, and X_C, respectively. Provided X_A and X_B are independent given X_C ($A \perp B \mid C$) then we have for all $M_A \in \sigma(X_A)$, $M_B \in \sigma(X_B)$, and $M_C \in \sigma(X_C)$ with $P(X_C \in M_C) > 0$:

$$P(X_A \in M_A, X_B \in M_B \mid X_C \in M_C) = P(X_A \in M_A \mid X_C \in M_C) P(X_B \in M_B \mid X_C \in M_C).$$

In particular, this is the case for $M_A = (-\infty, x_A]$, $M_B = (-\infty, x_B]$, and $M_C = (-\infty, x_C]$ with $x_A \in I\!R^{|X_A|}, x_B \in I\!R^{|X_B|}$, and $x_C \in I\!R^{|X_C|}$ arbitrary, but fixed. This yields

$$\frac{P(X_A \leq x_A, X_B \leq x_B, X_C \leq x_C)}{P(X_C \leq x_C)} = \frac{P(X_A \leq x_A, X_C \leq x_C)}{P(X_C \leq x_C)} \cdot \frac{P(X_B \leq x_B, X_C \leq x_C)}{P(X_C \leq x_C)}$$

and implies for all $x \in I\!\!R^p$ with $F_C(x_C) \neq 0$ the following factorization formula for cdf's which is analogous to Equation (1.6)

$$F(x) = F_{A\cup B\cup C}(x_A, x_B, x_C) = \frac{F_{A\cup C}(x_A, x_C) F_{B\cup C}(x_B, x_C)}{F_C(x_C)}. \tag{2.7}$$

In addition to the factorization of the joint cdf into marginal cdf's, it is possible to show that the joint cdf decomposes into a product of marginal and conditional cdf's analogously to the well–known result for density functions. To derive this result, the following proposition is needed.

Proposition 2.1 *Let $X = X_V$ be a vector of random variables and assume that $A \perp B \mid C$ holds for a partition A, B, and C of V. Then we have for all $x \in I\!\!R^p$ with $F_C(x_C) \neq 0$:*

$$\frac{\left\{\frac{\partial}{\partial x_A} F_{A\cup C}(x_A, x_C)\right\} f_C(x_C)}{f_{A\cup C}(x_A, x_C) F_C(x_C)} = \frac{\left\{\frac{\partial}{\partial x_B} F_{B\cup C}(x_B, x_C)\right\} f_C(x_C)}{f_{B\cup C}(x_B, x_C) F_C(x_C)} = 1. \tag{2.8}$$

Proof:
From $A \perp B \mid C$ it follows that (1.6) holds and with Equation (2.7) the joint density $f(x)$ can be written as

$$\begin{aligned}
f(x) &= f_{A\cup B\cup C}(x_A, x_B, x_C) = \frac{\partial^3}{\partial x_A \partial x_B \partial x_C} F_{A\cup B\cup C}(x_A, x_B, x_C) \\
&= \frac{\partial^3}{\partial x_A \partial x_B \partial x_C} \frac{F_{A\cup C}(x_A, x_C) F_{B\cup C}(x_B, x_C)}{F_C(x_C)} \\
&= \frac{1}{F_C^2(x_C)} \frac{\partial^2}{\partial x_A \partial x_B} \Bigg(\left\{ \frac{\partial}{\partial x_C} F_{A\cup C}(x_A, x_C) F_{B\cup C}(x_B, x_C) \right\} F_C(x_C) \\
&\qquad - F_{A\cup C}(x_A, x_C) F_{B\cup C}(x_B, x_C) f_C(x_C) \Bigg) \\
&= \frac{1}{F_C^2(x_C)} \frac{\partial^2}{\partial x_A \partial x_B} \Bigg(\left\{ \frac{\partial}{\partial x_C} F_{A\cup C}(x_A, x_C) \right\} F_{B\cup C}(x_B, x_C) F_C(x_C) \\
&\qquad + F_{A\cup C}(x_A, x_C) \left\{ \frac{\partial}{\partial x_C} F_{B\cup C}(x_B, x_C) \right\} F_C(x_C) \\
&\qquad - F_{A\cup C}(x_A, x_C) F_{B\cup C}(x_B, x_C) f_C(x_C) \Bigg) \\
&= \frac{1}{F_C^2(x_C)} \frac{\partial}{\partial x_A} \Bigg(\left\{ \frac{\partial}{\partial x_C} F_{A\cup C}(x_A, x_C) \right\} \left\{ \frac{\partial}{\partial x_B} F_{B\cup C}(x_B, x_C) \right\} F_C(x_C) \\
&\qquad + F_{A\cup C}(x_A, x_C) f_{B\cup C}(x_B, x_C) F_C(x_C) \\
&\qquad - F_{A\cup C}(x_A, x_C) \left\{ \frac{\partial}{\partial x_B} F_{B\cup C}(x_B, x_C) \right\} f_C(x_C) \Bigg) \\
&= \frac{1}{F_C^2(x_C)} \Bigg(f_{A\cup C}(x_A, x_C) \left\{ \frac{\partial}{\partial x_B} F_{B\cup C}(x_B, x_C) \right\} F_C(x_C) \\
&\qquad + \left\{ \frac{\partial}{\partial x_A} F_{A\cup C}(x_A, x_C) \right\} f_{B\cup C}(x_B, x_C) F_C(x_C)
\end{aligned}$$

$$
\begin{aligned}
&\qquad - \left\{ \frac{\partial}{\partial x_A} F_{A \cup C}(x_A, x_C) \right\} \left\{ \frac{\partial}{\partial x_B} F_{B \cup C}(x_B, x_C) \right\} f_C(x_C) \Big) \\
= &\ \frac{f_{A \cup C}(x_A, x_C) f_{B \cup C}(x_B, x_C)}{f_C(x_C)} \\
&\cdot \Big(\frac{\left\{ \frac{\partial}{\partial x_B} F_{B \cup C}(x_B, x_C) \right\} f_C(x_C)}{f_{B \cup C}(x_B, x_C) F_C(x_C)} + \frac{\left\{ \frac{\partial}{\partial x_A} F_{A \cup C}(x_A, x_C) \right\} f_C(x_C)}{f_{A \cup C}(x_A, x_C) F_C(x_C)} \\
&\qquad - \frac{\left\{ \frac{\partial}{\partial x_A} F_{A \cup C}(x_A, x_C) \right\} \left\{ \frac{\partial}{\partial x_B} F_{B \cup C}(x_B, x_C) \right\} f_C^2(x_C)}{f_{A \cup C}(x_A, x_C) f_{B \cup C}(x_B, x_C) F_C^2(x_C)} \Big).
\end{aligned}
$$

Using the abbreviations

$$
a = \frac{\left\{ \frac{\partial}{\partial x_A} F_{A \cup C}(x_A, x_C) \right\} f_C(x_C)}{f_{A \cup C}(x_A, x_C) F_C(x_C)} \quad \text{and} \quad b = \frac{\left\{ \frac{\partial}{\partial x_B} F_{B \cup C}(x_B, x_C) \right\} f_C(x_C)}{f_{B \cup C}(x_B, x_C) F_C(x_C)}
$$

it follows that $a + b - ab = 1$ or that $(1 - a)(1 - b) = 0$ which is fulfilled for $a = 1$ or $b = 1$. Since the two factors a and b have an identical structure it can be concluded that $a = b = 1$. □

From Equation (2.8) it is easily seen that

$$
f_{B|C}(x_B|x_C) = \frac{f_{B \cup C}(x_B, x_C)}{f_C(x_C)} = \frac{\frac{\partial}{\partial x_B} F_{B \cup C}(x_B, x_C)}{F_C(x_C)}.
$$

Therefore, we have

$$
F_{B|C}(x_B|x_C) = \int_{-\infty}^{x_B} f_{B|C}(u|x_C) du = \int_{-\infty}^{x_B} \frac{\frac{\partial}{\partial u} F_{B \cup C}(u, x_C)}{F_C(x_C)} du = \frac{F_{B \cup C}(x_B, x_C)}{F_C(x_C)}.
$$

Together with Equation (2.7), this result leads immediatly to a factorization formula which is dual to Equation (1.5) for the case that $A \perp B \mid C$ holds:

$$
F(x) = F_{A \cup C}(x_A, x_C) F_{B|C}(x_B|x_C). \tag{2.9}
$$

In the following, it is shown that the factorization of certain conditional probabilities is sufficient for conditional independence. The benefit of this statement is twofold. On the one hand, in some cases it is simpler to prove the factorization for some probabilities than for the joint density or cdf. On the other hand, this is the missing argument for the inversion of the above results, i.e. the conclusion from the factorization of the cdf given by Equations (2.7) and (2.9) to the conditional independence of A and B given C.

Proposition 2.2 *Let $U = (X^T, Y^T, Z^T)^T$ be a vector of random variables. The vectors X and Y are conditionally independent given Z if for all $z \in \mathbb{R}^{|Z|}$ with $P(Z \leq z) > 0$ the following holds*

$$
P(X \leq x, Y \leq y \mid Z \leq z) = P(X \leq x \mid Z \leq z) P(Y \leq y \mid Z \leq z). \tag{2.10}
$$

Proof:
The conclusion from conditional independence to the above factorization of the conditional probability is directly obtained from the definition of the conditional independence of X and Y given Z. Here, the inversion is to be shown which is the more interesting part, i.e. that the independence of X and Y given Z follows from (2.10). Note that the conditioning is a set, namely the set $\{Z \leq z\}$ and not the point $Z = z$. As defined, X and Y are conditionally independent given Z if for all $z \in I\!R^{|Z|}$ with $f(z) > 0$ it holds that

$$F_{X,Y|Z}(x, y|z) = F_{X|Z}(x|z)F_{Y|Z}(y|z). \tag{2.11}$$

The cdf of the conditional distribution of X given $Z = z$ is defined as

$$\begin{aligned}
F_{X|Z}(x|z) &= \lim_{h \searrow 0} P(X \leq x \mid z - h < Z \leq z) \\
&= \lim_{h \searrow 0} \frac{P(X \leq x, z - h < Z \leq z)}{P(z - h < Z \leq z)} \\
&= \frac{\lim_{h \searrow 0} \frac{1}{h} P(X \leq x, z - h < Z \leq z)}{\lim_{h \searrow 0} \frac{1}{h} P(z - h < Z \leq z)} \\
&= \frac{\int_{-\infty}^{x} f_{X,Z}(t, z) dt}{f_Z(z)}.
\end{aligned}$$

Note that $F_{Y|Z}$ and $F_{X,Y|Z}$ can be written analogously. Let $F(x, y, z)$ denote the cdf of $U = (X^T, Y^T, Z^T)^T$, $F_{X,Z}(x, z)$, $F_{Y,Z}(y, z)$, and $F_Z(z)$ the cdf's of the corresponding marginal distributions of $(X^T, Z^T)^T$, $(Y^T, Z^T)^T$, and Z. In the following Equation (2.10) is used rewritten as

$$F(x, y, z)F_Z(z) = F_{X,Z}(x, z)F_{Y,Z}(y, z). \tag{2.12}$$

With these notations and since all involved limits exist we get from Equation (2.10)

$$\begin{aligned}
& \lim_{h \searrow 0} P(X \leq x \mid z - h < Z \leq z) \lim_{h \searrow 0} P(Y \leq y \mid z - h < Z \leq z) \\
= \; & \lim_{h \searrow 0} P(X \leq x, Y \leq y \mid z - h < Z \leq z) \\
\Leftrightarrow \; & \frac{\lim_{h \searrow 0} \frac{1}{h} P(X \leq x, z - h < Z \leq z)}{\lim_{h \searrow 0} \frac{1}{h} P(z - h < Z \leq z)} \frac{\lim_{h \searrow 0} \frac{1}{h} P(Y \leq y, z - h < Z \leq z)}{\lim_{h \searrow 0} \frac{1}{h} P(z - h < Z \leq z)} \\
= \; & \frac{\lim_{h \searrow 0} \frac{1}{h} P(X \leq x, Y \leq y, z - h < Z \leq z)}{\lim_{h \searrow 0} \frac{1}{h} P(z - h < Z \leq z)} \\
\Leftrightarrow \; & \lim_{h \searrow 0} \frac{1}{h} P(X \leq x, z - h < Z \leq z) \lim_{h \searrow 0} \frac{1}{h} P(Y \leq y, z - h < Z \leq z) \\
= \; & \lim_{h \searrow 0} \frac{1}{h} P(X \leq x, Y \leq y, z - h < Z \leq z) \lim_{h \searrow 0} \frac{1}{h} P(z - h < Z \leq z) \\
\Leftrightarrow \; & \lim_{h \searrow 0} \{ P(X \leq x, z - h < Z \leq z) P(Y \leq y, z - h < Z \leq z) \\
& \quad - P(X \leq x, Y \leq y, z - h < Z \leq z) P(z - h < Z \leq z) \} \\
= \; & 0.
\end{aligned}$$

The right–hand side can be written as

$$
\begin{aligned}
& \lim_{h\searrow 0}\{P(X\le x, z-h<Z\le z)P(Y\le y, z-h<Z\le z) \\
& \quad -P(X\le x, Y\le y, z-h<Z\le z)P(z-h<Z\le z)\} \\
= & \lim_{h\searrow 0}\{[P(X\le x, Z\le z)-P(X\le x, Z\le z-h)][P(Y\le y, Z\le z)-P(Y\le y, Z\le z-y)] \\
& \quad -\{[P(X\le x, Y\le y, Z\le z)-P(X\le x, Y\le y, Z\le z-h)] \\
& \qquad [P(Z\le z)-P(Z\le z-h)]\}\} \\
= & \lim_{h\searrow 0}\{[F_{X,Z}(x,z)-F_{X,Z}(x,z-h)][F_{Y,Z}(y,z)-F_{Y,Z}(y,z-y)] \\
& \quad -\{[F(x,y,z)-F(x,y,z-h)][F_Z(z)-F_Z(z-h)]\}\} \\
= & \lim_{h\searrow 0}\{F_{X,Z}(x,z)F_{Y,Z}(y,z)-F_{X,Z}(x,z)F_{Y,Z}(y,z-h) \\
& \quad -F_{X,Z}(x,z-h)F_{Y,Z}(y,z)+F_{X,Z}(x,z-h)F_{Y,Z}(y,z-h) \\
& \quad -F(x,y,z)F_Z(z)+F(x,y,z)F_Z(z-h) \\
& \quad +F(x,y,z-h)F_Z(z)-F(x,y,z-h)F_Z(z-h)\} \\
= & \lim_{h\searrow 0}\{F(x,y,z)F_Z(z)-F_{X,Z}(x,z)F_{Y,Z}(y,z-h) \\
& \quad -F_{X,Z}(x,z-h)F_{Y,Z}(y,z)+F(x,y,z-h)F_Z(z-h) \\
& \quad -F(x,y,z)F_Z(z)+F(x,y,z)F_Z(z-h) \\
& \quad +F(x,y,z-h)F_Z(z)-F(x,y,z-h)F_Z(z-h)\} \\
= & \lim_{h\searrow 0}\{-F_{X,Z}(x,z)F_{Y,Z}(y,z-h)-F_{X,Z}(x,z-h)F_{Y,Z}(y,z) \\
& \quad +F(x,y,z)F_Z(z-h)+F(x,y,z-h)F_Z(z)\},
\end{aligned}
$$

where in the fourth step Equation (2.12) is applied to the first and the fourth factor. Finally, this leads to

$$
\begin{aligned}
& \lim_{h\searrow 0}\{-F_{X,Z}(x,z)F_{Y,Z}(y,z-h)-F_{X,Z}(x,z-h)F_{Y,Z}(y,z) \\
& \qquad +F(x,y,z)F_Z(z-h)+F(x,y,z-h)F_Z(z)\} \\
= & -F_{X,Z}(x,z)F_{Y,Z}(y,z)-F_{X,Z}(x,z)F_{Y,Z}(y,z)+F(x,y,z)F_Z(z)+F(x,y,z)F_Z(z) \\
= & -2F(x,y,z)F_Z(z)+2F(x,y,z)F_Z(z)=0,
\end{aligned}
$$

where again Equation (2.12) is used. □

In the next section, it is demonstrated how the above results can be used in the framework of ML estimation in graphical models. This is for example necessary for the family of KS distributions which is introduced in the discussion. Especially, the argument given by proposition 2.2 is very helpful when working with this distributions.

3 An application: decomposition of ML estimation in graphical models

As mentioned in the introduction, a graphical model has the property that independence statements can be read off the graph. Another benefit of the graphical representation of the association structure among the variables concerns the possible simplification of the ML estimation. For special cases it has been shown that an appropriate decomposition of the graph results in a decomposition of the estimation problem into smaller ones each having a reduced number of parameters to be estimated and belonging to graphical models based on subgraphs of the original graph. A detailed discussion of these problems in case of Conditional–Gaussian distribution can be found in Frydenberg and Lauritzen (1989) and Frydenberg (1990).
The decomposition of the ML estimation is possible whenever the likelihood function factorizes into functions that are distinct with respect to the unknown parameter vector, say ω. Thus, the maximization reduces to the separate maximization of each factor. This procedure is justified by a general concept proposed by Barndorff–Nielsen (1978). The idea is to show that the investigated family of distributions $\mathcal{P}$ can be written as a product space $\mathcal{P}_T \times \mathcal{P}^T$ for a statistic $T = T(X)$, where $\mathcal{P}_T$ denotes the set of distributions of T and $\mathcal{P}^T$ the set of conditional distributions of X given T. This implies on the one hand that any density $p \in \mathcal{P}$ factorizes into the product

$$p(x;\omega) = p_T(T(x);\omega_1)\, p^T(x;\omega_2|T(X) = T(x)) \tag{3.13}$$

for $p_T \in \mathcal{P}_T$ and $p^T \in \mathcal{P}^T$. On the other hand, it holds that any product (3.13) for arbitrary elements $p_T \in \mathcal{P}_T$ and $p^T \in \mathcal{P}^T$ is an element of $\mathcal{P}$. This defines the statistic T as a cut in $\mathcal{P}$ (Barndorff–Nielsen, 1978, p. 50). This argument is applied to a general graphical model $\mathcal{M}(\mathcal{G})$, i.e. $\mathcal{M}(\mathcal{G})$ is the set of all distributions P of a family of distributions $\tilde{\mathcal{P}}$ fulfilling the Markov property of a given graph $\mathcal{G} = (V, E)$. In the following, we identify elements $P \in \mathcal{M}(\mathcal{G})$ with their density and their cdf, respectively, i.e., $f(\cdot) \in \mathcal{M}(\mathcal{G})$ and $F(\cdot) \in \mathcal{M}(\mathcal{G})$ refer to the same element of $\mathcal{M}(\mathcal{G})$.
For a graph $\mathcal{G} = (V, E)$ and $A \subset V$ we define the set of marginal distributions as $\mathcal{M}(\mathcal{G})_A = \{f_A(\cdot) \mid f(\cdot) \in \mathcal{M}(\mathcal{G})\}$ and the set of conditional distributions of $\mathcal{M}(\mathcal{G})$ given X_A as $\mathcal{M}(\mathcal{G})^A = \{f_{V\setminus A|A}(\cdot|\cdot) \mid f(\cdot) \in \mathcal{M}(\mathcal{G})\}$. These notations are also used to denote sets like

$$(\mathcal{M}(\mathcal{G})_{B\cup C})^C = \{f_{B|C}(\cdot|\cdot) \mid f(\cdot) \in \mathcal{M}(\mathcal{G})_{B\cup C}\}$$

for disjoint subsets B, C of V. Note that this differs from

$$\mathcal{M}(\mathcal{G}_{B\cup C})^C = \{f_{B|C}(\cdot|\cdot) \mid f(\cdot) \in \mathcal{M}(\mathcal{G}_{B\cup C})\},$$

insofar as $\mathcal{M}(\mathcal{G}_{B\cup C})$ denotes a graphical model corresponding to the graph $\mathcal{G}_{B\cup C}$ whereas $\mathcal{M}(\mathcal{G})_{B\cup C}$ is the family of marginal distributions of a graphical model corresponding to the graph $\mathcal{G}$. This difference is essential for the following argumentation. Within the scope of graphical models, the idea of a cut is used in a slightly modified way. For a partition A, B, and C of V the statistic $T(x) = x_{A\cup C}$ is said to be a cut in $\mathcal{M}(\mathcal{G})$ if the following three conditions hold:

(i) $\mathcal{M}(\mathcal{G})_{A\cup C} = \mathcal{M}(\mathcal{G}_{A\cup C})$,

(ii) $\mathcal{M}(\mathcal{G})^{A\cup C} = \mathcal{M}(\mathcal{G}_{B\cup C})^C$,

(iii) $\mathcal{M}(\mathcal{G}) = \mathcal{M}(\mathcal{G}_{A\cup C}) \times \mathcal{M}(\mathcal{G}_{B\cup C})^C$.

This definition of a cut is more restrictive than the one of Barndorff–Nielsen described above as two closure properties are called for besides the factorization criterion which shows up again in condition (iii). These additional properties (i) and (ii) guarantee that the factors of the product space are related to graphical models corresponding to subgraphs. For instance, condition (i) equals the so–called collapsibility of graphical models onto the set $A \cup C$ (cf. Frydenberg and Lauritzen, 1989) and describes that the family of marginal distributions $\mathcal{M}(\mathcal{G})_{A\cup C}$ of a graphical model $\mathcal{M}(\mathcal{G})$ coincides with the graphical model $\mathcal{M}(\mathcal{G}_{A\cup C})$ related with the subgraph $\mathcal{G}_{A\cup C}$ of $\mathcal{G}$.
In the present context, we are mainly interested in condition (iii). Therefore, we assume that (i) and (ii) hold and we discuss additional conditions for (iii) to be valid. Under the assumption $A \perp B \mid C$, the inclusion $\mathcal{M}(\mathcal{G}) \subseteq \mathcal{M}(\mathcal{G}_{A\cup C}) \times \mathcal{M}(\mathcal{G}_{B\cup C})^C$ follows directly from the factorization of the joint density (1.5) combined with (i) and (ii):

$$\mathcal{M}(\mathcal{G}) \subseteq \mathcal{M}(\mathcal{G})_{A\cup C} \times \mathcal{M}(\mathcal{G})^{A\cup C} = \mathcal{M}(\mathcal{G}_{A\cup C}) \times \mathcal{M}(\mathcal{G}_{B\cup C})^C.$$

The opposite direction, i.e.

$$\mathcal{M}(\mathcal{G}_{A\cup C}) \times \mathcal{M}(\mathcal{G}_{B\cup C})^C \subseteq \mathcal{M}(\mathcal{G}). \tag{3.14}$$

is in general more difficult to prove. For this purpose, we show that for arbitrary density functions $g_{A\cup C}(\cdot) \subseteq \mathcal{M}(\mathcal{G}_{A\cup C})$ and $h_{B|C}(\cdot) \subseteq \mathcal{M}(\mathcal{G}_{B\cup C})^C$ the product $g_{A\cup C}(x_A, x_C)h_{B|C}(x_B|x_C)$ is in $\mathcal{M}(\mathcal{G})$. In contrast to the above inclusion, this crucially depends on the properties of the underlying multivariate distribution especially on its factorization properties with respect to the parameter vector. That means, it must be checked whether the density of the assumed distribution fulfills the stated Condition (3.14) or not. In the following, it is shown that there is an equivalent condition for the cdf. We first need some further notations. Let $G_{A\cup C}(\cdot)$ denote the cdf of $g_{A\cup C}(\cdot)$ and for

$$h_{B|C}(x_B|x_C) = \frac{h_{B\cup C}(x_B, x_C)}{h_C(x_C)}$$

with $h_{B\cup C}(\cdot) \in \mathcal{M}(\mathcal{G}_{B\cup C})$ and marginal density $h_C(\cdot) \in \mathcal{M}(\mathcal{G}_{B\cup C})_C$, let $H_{B\cup C}(\cdot)$ and $H_C(\cdot)$ denote the corresponding cdf's. The general idea is to show that

$$\frac{G_{A\cup C}(x_A, x_C)H_{B\cup C}(x_B, x_C)}{H_C(x_C)} = G_{A\cup C}(x_A, x_C)H_{B|C}(x_B|x_C) = F(x)$$

is the cdf of $g_{A\cup C}(x_A, x_C)h_{B|C}(x_B|x_C)$, i.e. that $f(\cdot) \in \mathcal{M}(\mathcal{G})$ is equivalent to $F(\cdot) \in \mathcal{M}(\mathcal{G})$. Therefore, define $H^*(x)$ as

$$H^*(x) = H_{B\cup C}(x_{B\cup C}) \cdot \prod_{i\in A} F_i(x_i)$$

with corresponding density function $h^*(\cdot)$. Then it trivially holds that $H^*(\cdot) \in \mathcal{M}(\mathcal{G})$. Since $A \perp B \mid C$ is assumed to hold for all elements of $\mathcal{M}(\mathcal{G})$, Lemma 2.1 yields

$$1 = \frac{\left\{\frac{\partial}{\partial x_B} H^*_{B\cup C}(x_B, x_C)\right\} h^*_C(x_C)}{h^*_{B\cup C}(x_B, x_C)H^*_C(x_C)} = \frac{\left\{\frac{\partial}{\partial x_B} H_{B\cup C}(x_B, x_C)\right\} h_C(x_C)}{h_{B\cup C}(x_B, x_C)H_C(x_C)}.$$

In analogy to the proof of Lemma 2.1, the density $f(\cdot)$ of $F(\cdot)$ can be derived as

$$\begin{aligned}
\frac{\partial}{\partial x}F(x) &= \frac{\partial^3}{\partial x_A \partial x_B \partial x_C}\frac{G_{A\cup C}(x_A, x_C)H_{B\cup C}(x_B, x_C)}{H_C(x_C)} \\
&= \frac{g_{A\cup C}(x_A, x_C)h_{B\cup C}(x_B, x_C)}{h_C(x_C)} \\
&\quad \cdot\Bigg(\frac{\left\{\frac{\partial}{\partial x_B} H_{B\cup C}(x_B, x_C)\right\} h_C(x_C)}{h_{B\cup C}(x_B, x_C)H_C(x_C)} + \frac{\left\{\frac{\partial}{\partial x_A} G_{A\cup C}(x_A, x_C)\right\} h_C(x_C)}{g_{A\cup C}(x_A, x_C)H_C(x_C)} \\
&\quad - \frac{\left\{\frac{\partial}{\partial x_B} H_{B\cup C}(x_B, x_C)\right\} h_C(x_C)}{h_{B\cup C}(x_B, x_C)H_C(x_C)} \cdot \frac{\left\{\frac{\partial}{\partial x_A} G_{A\cup C}(x_A, x_C)\right\} h_C(x_C)}{g_{A\cup C}(x_A, x_C)H_C(x_C)}\Bigg) \\
&= \frac{g_{A\cup C}(x_A, x_C)h_{B\cup C}(x_B, x_C)}{h_C(x_C)} = g_{A\cup C}(x_A, x_C)h_{B|C}(x_B|x_C) = f(x).
\end{aligned}$$

This implies that the condition $f(x) = g_{A\cup C}(\cdot)h_{B|C}(\cdot|\cdot) \in \mathcal{M}(\mathcal{G})$ is equivalent to $F(\cdot) = G_{A\cup C}(\cdot)H_{B|C}(\cdot|\cdot) \in \mathcal{M}(\mathcal{G})$ which is expressed in the following corollary.

Corollary 3.1 *Let $X = X_V$ be a vector of random variables and A, B, and C a partition of V with $A \perp B \mid C$. Using the above definitions it holds that*

$$f(\cdot) = g_{A\cup C}(\cdot)h_{B|C}(\cdot|\cdot) \in \mathcal{M}(\mathcal{G}) \Leftrightarrow F(\cdot) = G_{A\cup C}(\cdot)H_{B|C}(\cdot|\cdot) \in \mathcal{M}(\mathcal{G}).$$

Summarizing, it can be concluded that the factorization formulae for the cdf of Section 2 allow to switch over from a condition for the density function to a condition for the cdf which may simplify the calculations in special cases. In the following section, a situation is described where the problem occurs that it is impossible to check condition $f(\cdot) \in \mathcal{M}(\mathcal{G})$ whereas the discussion of condition $F(\cdot) \in \mathcal{M}(\mathcal{G})$ is successful.

4 Discussion

As already mentioned, the above results may be of importance when distributions other than the Conditional–Gaussian distribution are assumed as joint distribution for X_V in graphical models. For instance, Koehler and Symanowski (1995) introduce a class of multivariate distribution families which allows to model on the one hand complex associations among arbitrary subsets of the variable set and on the other hand pairwise independences in the margins. In addition, this distribution family fulfills the equivalence of the different Markov properties which is a minimal requirement for a model distribution as far as graphical models are concerned. All these properties together state that this distribution family is worth being discussed in the framework of graphical models.

With $V = \{1, \ldots, p\}$ and $\mathcal{V}$ being the powerset of V, let $X = X_V = (X_1, \ldots, X_p)^T$ denote a vector of random variables with marginal cdf's $F_i(\cdot)$, $i \in V$. The joint distribution of X is assumed to be given by the following cdf

$$F(x_1, \ldots, x_p) = \prod_{i \in V} F_i(x_i) \prod_{I \in \mathcal{I}} c_I(x)^{-\alpha_I}. \tag{4.15}$$

For all sets $I \in \mathcal{I} = \{I \in \mathcal{V} \text{ with } |I| \geq 2\}$ let $\mathbb{R} \ni \alpha_I \geq 0$ and for all $i \in V$ let $\mathbb{R} \ni \alpha_i > 0$ with $\alpha_{i+} = \sum_{I \in \mathcal{V}, i \in I} \alpha_I < \infty$. For all $I \in \mathcal{I}$ the factors $c_I(x)$ in Equation (4.15) are defined as

$$c_I(x) = \sum_{i \in I} \{ \prod_{\substack{j \in I \\ j \neq i}} u_j(x_j) \} - (|I| - 1) \prod_{i \in I} u_i(x_i)$$

with $u_i(x_i) = F_i(x_i)^{\frac{1}{\alpha_{i+}}}$ for all $i \in V$. Here, the structure of the cdf is fairly easy. It factorizes into the product of the marginal cdf's and a product of association terms. These distributions can be constructed for almost any given univariate marginal distributions by adding interaction terms and can be viewed as a generalization of the generalized Burr–Pareto–logistic distributions.

If we assume that marginal density functions $f_i(\cdot)$ exist for all $i \in V$ it can easily be shown that the joint density function also exists (Koehler and Symanowski, 1995). However, in contrast to the cdf the functional representation of the density function is rather complicated. Besides the product of the marginal densities there are more complex factors with additive components due to the derivation. This leads to a formula which is not easy to handle and therefore, an analysis of factorization properties already fails in the simplest situations. In this case, the indirect argument via cdf's using the factorization formula (2.9) and Proposition 2.2 allows to discuss the problem described above and leads to the conditions which are needed to obtain a decomposition and therefore a simplification of the ML estimation.

Another possible application is the construction of the joint multivariate distribution from given conditional and marginal distributions in situations where the joint density does not exist or is not of primary interest. Given a certain association structure among the involved variables, i.e. certain conditional independences, the joint multivariate distribution can be derived making use of Equation (2.9) in analogy to Bayes formula. A more detailed examination of this idea and related problems could be a rewarding field for further research.

Acknowledgment
Financial support by the SFB 386, Deutsche Forschungsgemeinschaft is gratefully acknowledged.

References

Barndorff–Nielsen, O.E. (1978). *Information and exponential families in statistical theory*. John Wiley and Sons, New York.

Dawid, A.P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society B*, 41, 1–31.

Dawid, A.P. (1980). Conditional independence for statistical operations. *The Annals of Statistics*, 8, 598–617.

Frydenberg, M. (1990). Marginalization and collapsibility in graphical interaction models. *The Annals of Statistics*, 18, 790–805.

Frydenberg, M. and Lauritzen, S.L. (1989). Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika*, 76, 539–555.

Koehler, K.J. and Symanowski, J.T. (1995). Constructing multivariate distributions with specific marginal distributions. *Journal of Multivariate Analysis*, 55, 261–282.

Lauritzen, S.L. (1996). *Graphical models*. Oxford University Press, Oxford.

Lauritzen, S.L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17, 31–57.

Wermuth, N. and Lauritzen, S.L. (1990). On substantive research hypotheses, conditional independence, graphs and graphical chain models. *Journal of the Royal Statistical Society B*, 52, 21–50.