ДОПОВІДІ
НАЦІОНАЛЬНОЇ
АКАДЕМІЇ НАУК
УКРАЇНИ

12 • 2016

*ІНФОРМАТИКА*

doi: https://doi.org/10.15407/dopovidi2016.12.017

УДК 004.855:519.216

**О.С. Балабанов**

Інститут програмних систем НАН України, Київ
E-mail: bas@isofts.kiev.ua

# Про характерні співвідношення кореляцій в деяких системах лінійних структуральних рівнянь

*(Представлено академіком НАН України П. І. Андоном)*

*Для ймовірнісної лінійної моделі з циклічною структурою із чотирма змінними знайдено і доведено два простих обмеження типу нерівність на наборі кореляцій. Кожне з цих обмежень (що включає дві та три кореляції відповідно) дає змогу спростувати базову модель на користь альтернативної моделі, яка відрізняється додатковим "діагональним" зв'язком.*

***Ключові слова:*** *структура моделі, лінійні залежності, кореляція, обмеження типу нерівність, прихована змінна, верифікація моделі, структура зв'язків, марковська властивість.*

Відомо, що структура системи залежностей накладає на характеристики моделі певні обмеження [1]. Ці бмеження можуть бути використані для верифікації моделі навіть в умовах неповної спостережуваності (ймовірнісної невизначеності). Марковські властивості моделей виражаються як обмеження типу рівність. Коли дві чи більше змінних моделі зазнають впливу деякої прихованої (латентної) змінної, неможливо безпосередньо тестувати відповідну марковську властивість. Але в деяких моделях з прихованою змінною можна тестувати імплікації марковських властивостей. Наприклад, коли прихованою є вузлова ("центральна") змінна в деревовидній структурі, то (в певних класах моделей) виконуються прості обмеження типу рівність для набору парних залежностей [1, 2]. В інших ситуаціях діють обмеження тільки типу нерівність. Славнозвісним прикладом є обмеження на кореляцію, встановлене теоремою Дж.С. Белла для відповідної квантової системи [3].

Найпростішим прикладом, на якому можна проілюструвати обмеження типу нерівність, є ймовірнісна модель із структурою ланцюга $X \to Q \to Y$. (Ця модель є більш загальною за звичайний марковський ланцюг і може включати змінні різних типів та розмірності). Аналітику може бути недоступним сумісне розподілення ймовірностей трьох змінних. Тоді можна аналізувати співвідношення парних залежностей. Вказана структура імплікує наступні нерівності для взаємної інформації (за Шенноном): Info $(X, Q) \geqslant$ Info $(X, Y)$, Info $(Q, Y) \geqslant$ Info $(X, Y)$. Коли такий ланцюг утворений з лінійних залежностей (а розподіли

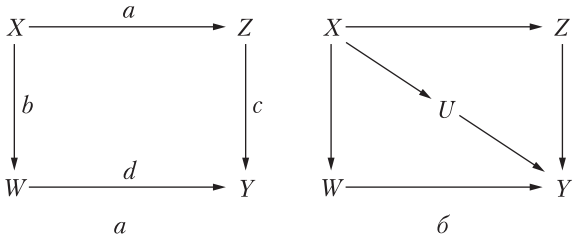


***Рис. 1***

ймовірностей — нормальні), отримуємо (більш сильне) обмеження типу рівність для коефіцієнтів кореляції: $\rho_{xy} = \rho_{xq}\,\rho_{qy}$. Моделі з циклічною структурою характеризувати значно важче. Спеціалістам з каузального моделювання відоме обмеження Верма (Verma-constraint) для розподілів умовних ймовірностей [4]. Це обмеження характеризує модель, де чотири дискретні змінні з'єднані як ланцюг, але "друга" та "четверта" змінна зазнають впливу прихованої змінної $H$. Якщо з такої моделі усунути "третю" змінну (посередника між змінними, залежними від прихованої $H$), то отримаємо модель іструментальної змінної, де каузальний зв'язок є "сплутаним" (конфаундованим) латентною змінною. Для цієї моделі можна знайти тільки специфічне й доволі "непрозоре" обмеження типу нерівність [5] (причому необхідно, щоб "середня" змінна була дискретною). Далі, якщо остання змінна додатково має ще окремий фактор впливу, аналіз характеристик стає ще складнішим [6]. Стан досліджень цього напряму відображено в [7]. Для лінійних моделей з двома чи більше латентними факторами відомі обмеження типу рівність [8]. На жаль, з огляду на форму тих обмежень (поліном на основі коваріацій), перспектива побудови достатньо потужного тесту верифікації моделі — сумнівна.

В даній роботі розглядається ймовірнісна лінійна модель, структура якої — нетриангульований цикл з одним "колізором". Тобто аналізується наступна система структуральних рівнянь:

$$x = \varepsilon_1;\ z = ax + \varepsilon_2;\ w = bx + \varepsilon_3;\ y = cz + dw + \varepsilon_4, \tag{1}$$

де $\varepsilon_i \sim \mathrm{N}(m_i, \sigma_i^2)$ та $\varepsilon_i \perp \varepsilon_j \quad \forall i \neq j$.

Структура цієї моделі відображена на рис. 1, *а*.

Проблема може полягати в наступному. Структура зв'язків може бути відомою не достеменно (не вичерпно); наприклад, не виключено, що існує третій шлях впливу змінної $X$ на $Y$, як показано на рис. 1, *б*. Цей зв'язок може бути безпосереднім або діяти через змінну $U$ (тоді останнє рівняння системи (1) набуває вигляду $y = cz + dw + hu + \varepsilon_4$ й додається рівняння для $U$). Якщо змінна $U$ — прихована, дві відповідні кореляції будуть невідомі. Кореляції для деяких інших пар змінних теж можуть бути недоступні, і тоді неможливо обчислити частинні кореляції $\rho_{xy.zw}$ та $\rho_{xy.zwu}$. Питання: як в умовах неповної інформації верифікувати присутність або відсутність зв'язку $X$—$Y$ в структурі моделі (по "головній діагоналі")? До речі, відсутність (додаткового) безпосереднього зв'язку між $Z$ та $W$ можна верифікувати перевіркою $\rho_{zw} = \rho_{xz}\,\rho_{xw}$.

Оскільки в моделі (1) між $X$ та $Y$ існують два шляхи залежності, то (на відміну від ланцюгової моделі $X \to Z \to Y$) "непряма" кореляція $\rho_{xy}$ не завжди буде меншою за "реберні" кореляції $\rho_{xz}$ та $\rho_{zy}$. Але модель (1) імплікує деякі інші обмеження на відношення кореляцій. Зауважимо, що для кожного набору значень параметрів в (1) можна знайти інший набір, такий, що $\sigma_1 = 1$ й разом з тим всі кореляції залишаться незмінними. Отже, для аналізу системи кореляцій достатньо розглянути випадок $\sigma_1 = 1$.

**Твердження 1**. *В системі (1) за будь-яких значень параметрів виконується співвідношення* $\rho_{xy}^2 \leqslant \rho_{zy}^2$ *або співвідношення* $\rho_{xy}^2 \leqslant \rho_{wy}^2$.

**Доведення**. Нехай, від протилежного, не виконуються обидва співвідношення, тобто нехай буде $\rho^2_{xy} / \rho^2_{zy} > 1$ та $\rho^2_{xy} / \rho^2_{wy} > 1$.

Для моделі (1) з уточненням $\sigma_1 = 1$ маємо

$$\rho^2_{xy} / \rho^2_{zy} = var(z) cov^2(x, y) / cov^2(z, y) = (a^2 + \sigma_2^2) \frac{(ac + bd)^2}{(a(ac + bd) + c\sigma_2^2)^2},$$

$$\rho^2_{xy} / \rho^2_{wy} = var(w) cov^2(x, y) / cov^2(w, y) = (b^2 + \sigma_3^2) \frac{(ac + bd)^2}{(b(ac + bd) + d\sigma_3^2)^2}.$$

Прийнявши припущення $\rho^2_{xy} / \rho^2_{zy} > 1$, після тотожних алгебраїчних перетворень отримуємо

$$b^2 d^2 > a^2 c^2 + c^2 \sigma_2^2. \tag{2}$$

Аналогічно, з припущення $\rho^2_{xy} / \rho^2_{wy} > 1$ випливає

$$a^2 c^2 > b^2 d^2 + d^2 \sigma_3^2. \tag{3}$$

Підсумовуючи (2) та (3), отримуємо $0 > c^2 \sigma_2^2 + d^2 \sigma_3^2$. Очевидно, останнє задовольнити неможливо, отже, припущення від протилежного спростоване. Твердження доведене.

Зазначимо, що аналогічні співвідношення для перших ланок шляхів зв'язку $X$ з $Y$ (тобто для $\rho_{xz}$ та $\rho_{xw}$) в загальному випадку не чинні. Така "нерівноправність" спричинена тим, що кореляції $\rho_{xz}$ та $\rho_{xw}$ визначаються тільки локальними параметрами, а кореляції $\rho_{zy}$ та $\rho_{wy}$ враховують вклад всіх ребер (зв'язків) моделі. Це пояснюється позицією зв'язків в циклі структури моделі (див. рис.1). Але існує інше обмеження за участю $\rho_{xz}$ та $\rho_{xw}$, й воно формулюється наступним чином.

**Твердження 2**. *В системі (1) за всіх значень параметрів виконується відношення* $\rho^2_{xy} \leqslant \rho^2_{xz} + \rho^2_{xw}$.

**Доведення**. В системі (1) маємо

$$\rho^2_{xz} = a^2 / (a^2 + \sigma_2^2);\ \rho^2_{xw} = b^2 / (b^2 + \sigma_3^2);$$

$$\rho^2_{xy} = (ac + bd)^2 / ((ac + bd)^2 + c^2 \sigma_2^2 + d^2 \sigma_3^2 + \sigma_4^2).$$

Нехай, від протилежного, $\rho^2_{xy} > \rho^2_{xz} + \rho^2_{xw}$. З огляду на позитивність всіх елементів наведених формул, це рівнозначно нерівності

$$a^2(b^2 + \sigma_3^2) + b^2(a^2 + \sigma_2^2) < \frac{(a^2 + \sigma_2^2)(b^2 + \sigma_3^2)(ac + bd)^2}{(ac + bd)^2 + c^2 \sigma_2^2 + d^2 \sigma_3^2 + \sigma_4^2}.$$

Виконавши алгебраїчні перетворення і групуючи терми $b^2 c^2 \sigma_2^4 - 2abcd\sigma_2^2 \sigma_3^2 + a^2 d^2 \sigma_3^4$, дістаємо $(bc\sigma_2^2 - ad\sigma_3^2)^2 + \sum_i t_i^2 < 0$, де $t_i$ — відповідні добутки параметрів моделі. Оскільки цю нерівність задовольнити неможливо, твердження доведене.

Зрозуміло, коли маємо $\rho^2_{xz} + \rho^2_{xw} > 1$, твердження 2 перестає працювати.

Об'єднання двох тверджень дає наступний результат.

**Наслідок**. *В системі (1) за будь-яких значень параметрів виконується відношення* $2\rho^2_{xy} \leqslant \rho^2_{xz} + \rho^2_{xw} + \max\{\rho^2_{zy}, \rho^2_{wy}\}$.

Можна записати наслідок у спрощеній (але слабшій) формі $2\rho_{xy}^2 \leqslant \rho_{xz}^2 + \rho_{xw}^2 + \rho_{zy}^2 + \rho_{wy}^2$. Для порівняння зазначимо, що в лінійній моделі з структурою ланцюга $X \to Z \to W \to Y$ чинне обмеження $4\rho_{xy}^2 \leqslant \rho_{xz}^2 + \rho_{xw}^2 + \rho_{zy}^2 + \rho_{wy}^2$. А в лінійній моделі із структурою ланцюга $X \to Z \to Y$ чинне обмеження $2\rho_{xy}^2 \leqslant \rho_{xz}^2 + \rho_{zy}^2$.

Обмеження, встановлені твердженнями 1 та 2, є жорсткими (в рамках структури базової моделі можна забезпечити крайній випадок, тобто рівність). Водночас ці обмеження є доволі ефективним інструментом в тому сенсі, що багато альтернативних моделей порушують ці обмеження. Наприклад, нехай (невідома аналітику) генеративна модель має структуру як на рис.1, *б* і описується рівняннями:

$$\begin{aligned} &x = \varepsilon_1;\ z = 0{,}8\ x + \varepsilon_2,\ w = 0{,}8\ x + \varepsilon_3; \\ &u = 2x + \varepsilon_5,\ y = z + w + 2u + \varepsilon_4, \end{aligned} \tag{4}$$

де всі $e_i$ взаємонезалежні, а їхні дисперсії дорівнюють 1. Тоді отримуємо $\rho_{xy}^2 = 0{,}818$; $\rho_{xz}^2 = 0{,}39$; $\rho_{xw}^2 = 0{,}39$; $\rho_{zy}^2 = 0{,}477$; $\rho_{wy}^2 = 0{,}477$. Як бачимо, обидва встановлених обмеження брутально порушуються. Базова модель спростовується.

Щоб мати можливість застосувати вказані критерії для верифікації основної моделі (1), необхідно знати принаймні три кореляції, а саме: $\rho_{xy}$, $\rho_{xz}$, $\rho_{xw}$ або $\rho_{xy}$, $\rho_{zy}$, $\rho_{wy}$. Завдяки особливій простоті запропонованих критеріїв можна побудувати ефективний статистичний тест верифікації моделі. Зрозуміло, що задовільнення вказаних обмежень не означає підтвердження базової моделі, як також не означає спростування нашої альтернативної моделі (чи будь-якої загальнішої). В такому разі обидві вказані моделі залишаються можливими. Для прикладу, змінимо в моделі (4) два коефіцієнти, так що два останніх рівняння набудуть вигляду $u = x + \varepsilon_5$, $y = z + w + 0{,}4\ u + \varepsilon_4$. В такій моделі обмеження з тверджень 1 та 2 не порушуються. Інше призначення встановлених обмежень — отримання верхніх або нижніх оцінок відповідних кореляцій для моделі заданої структури.

ЦИТОВАНА ЛІТЕРАТУРА


1. *Scheines R., Spirtes P., Glymour C.,* Meek C., Richardson T. The TETRAD project: Constraint based aids to causal model specification // Multivariate Behavioral Research. — 1998. — **33**, N 1. — P. 65—118.
2. *Андон П. І., Балабанов О. С.* До відкриття латентного бінарного фактора в статистичних даних категорного типа // Доп. НАН України. — 2008. — № 9. — С. 37–43.
3. *Bell J. S.* On the Einstein-Podolsky-Rosen paradox // Physics. — 1964. — **1**. — P. 195—200.
4. *Tian J., Pearl J.* On the testable implications of causal models with hidden variables // Proceed. of the 18th Conf. on Uncertainty in Artificial Intelligence (UAI-02). — San Francisco, CA: Morgan Kaufmann, 2002. — P. 519—527.
5. *Kang Ch., Tian J.* Inequality constraints in causal models with hidden variables // Proceed. of the 22nd Conf. on Uncertainty in Artificial Intelligence (UAI-06), Arlington, Virginia: AUAI Press, 2006, P. 233–240.
6. *Ramsahai R. R.* Causal bounds and observable constraints for non-deterministic models // J. Mach. Learn. Res. — 2012. — **13**. — P. 829—848.
7. *Evans R. J.* Graphical methods for inequality constraints in marginalized DAGs // 22nd Workshop on Machine Learning and Signal Processing / Santander, Spain, 2012, — P. 1—6. (see also: Preprint: arXiv:1209.2978v1 [math.ST]).
8. *Drton M., Sturmfels B., Sullivant S.* Algebraic factor analysis: tetrads, pentads and beyond // Probability Theory and Related Fields. — 2007. — **138,** N 3—4. — P. 463—493.

REFERENCES

1. *Scheines R., Spirtes P., Glymour C., Meek C., Richardson T.* Multivar. Behavioral Res., 1998, **33**, No 1: 65-118.
2. *Andon P. I., Balabanov O. S.* Dop. NAN Ukraine, 2008, No 9: 37-43 (in Ukrainian).
3. *Bell J. S.* Physics, 1964, **1**: 195-200.
4. *Tian J, Pearl J.* Proc. 18th Conf. on Uncertainty in Artif. Intellig., 2002: 519-527.
5. *Kang Ch, Tian J.* Proc. 22nd Conf. on Uncertainty in Artif. Intellig., 2006: 233-240.
6. *Ramsahai R. R.* J. Mach. Learn. Res. — 2012. — **13**: 829-848.
7. *Evans R. J.* 22nd Workshop on Machine Learning and Signal Processing, 2012, 1-6 (Preprint: arXiv:1209.2978v1 [math.ST]).
8. *Drton M., Sturmfels B., Sullivant S.* Probability Theory and Related Fields, 2007, **138**, No 3-4: 463-493.

*Надійшло до редакції 19.05.2016*

*А.С. Балабанов*

Институт программных систем НАН Украины, Киев
*E-mail:* bas@isofts.kiev.ua

О ХАРАКТЕРНЫХ СООТНОШЕНИЯХ
КОРРЕЛЯЦИЙ В НЕКОТОРЫХ СИСТЕМАХ
ЛИНЕЙНЫХ СТРУКТУРАЛЬНЫХ УРАВНЕНИЙ

*Для вероятностной линейной модели с цикличной структурой с четырьмя переменными найдены и доказаны два простых ограничения типа неравенства на наборе корреляций. Каждое из этих ограничений (включающее две и три корреляции соответственно) даёт возможность опровергнуть базовую модель в пользу альтернативной модели, которая отличается дополнительной "диагональной" связью.*

***Ключевые слова:*** *структура модели, линейные зависимости, корреляция, ограничение типа неравенство, скрытая переменная, верификация модели, структура связей, марковское свойство.*

*O.S. Balabanov*

Institute of Software Systems of the NAS of Ukraine, Kiev
*E-mail:* bas@isofts.kiev.ua

ON THE INTRINSIC RELATIONS OF CORRELATIONS
IN SOME SYSTEMS OF LINEAR STRUCTURAL EQUATIONS

*For a probabilistic linear model of cyclic structure with four variables, we prove two simple inequality-type constraints on the set of correlations. Each of the inequalities (comprising two and three correlations, respectively) facilitates the rejection of the basic model in favor of an alternative model, which differs in that it contains an additional "diagonal" connection.*

***Keywords:*** *model structure, linear dependences, correlation, inequality constraint, hidden variable, model verification, relationship structure, Markov property.*