EM или DA или EM и DA

гл. ас. Деян Лазаров,
Бургаски свободен университет

Представеното изследване е един вид продължение на анализа направен от автора в [8]. В цитираната студия се коментира, че появата на липсващи стоности (ЛС) при изследваните променливи в Наблюдението на работната сила в България през 2007 г. следват механизма на не случайно липсващи (НСЛ) стойности, което налога търсенето на модел описващ тяхната поява. След като се прилага процедура на кластъризация по медода К-средни (K-means) се достига до разделяне на данните на две: една по-малка група, наречена Кластър 7, със средни характеристики близки до 61 часа и една по-голяма група от останалите данни със средни характеристики при променливите с ЛС близки до 41 часа. Така възниква идеята за заместване ЛС, при механизма НСЛ, чрез третиране на данните в двете групи поотделно.

В настоящото изследване фокусът се насочва към сравнението на резултатите от прилагането на два концептуално много близки, но технически принадлежащи на два различни статистически подхода анализи. Единият анализ е **ЕМ алготитъмът**, при който като начин на постигане на резултатите се използва максимизирането на функцията на максималното правдоподобие (ФМП), а другият е **алгоритъмът за увеличаване на данните (Data augmentation algorithm)**, който за удобство се съкращава с DA[1], базиран на бейсовия подход и използващ информация от постериорните разпределения на променливите. И двата алгоритъма са итеративни и първият - ЕМ исторически е станал основа за създаването на вторият - DA. Въз основа на тези два алгоритъма, независимо един от друг, се въвеждат липсващите стойности в базата данни. На следваща стъпка се прави анализ на ЛС посредством комбинация от двата метода, а именно първоначалните параметри на разпределенията на променливите се взема от резултатите от ЕМ алгоритъма, а последващото въвеждане на ЛС се получава от DA алгоритъма. В резулат се направи опит за сравнение на трите подхода при рашаване на проблема с ЛС при наблюдението на работната сила в България (НРС) през 2007 г.

Наблюдението на работната сила в България - 2007 г.

В изследването са използват данни за цялата 2007 г. Наблюдавани са общо 160 признаци, част от които определят чрез демографски характеристики домакинството и единицата, лицето от домакинството, обект на изследване, друга част се отнасят за заетоста през последната седмица съответно на основна и допълнителна работа, както и форми на заетост и активност при търсене на работа и др. В анализа се налага редуциране на признаците поради няколко причини. От една страна това, че има преходи между въпросите в анкетната карта, т.е ако една единица даде определен отговор на даден въпрос то това води до автоматичното отпадане на въпроси в анкетната карта. От друга страна редица от признаците са технически идентификации на единиците в съвкупността, въведени от изследователския екип, които логически не се отнасят към изследваното явление, а именно липсващите стойности. Специфичен интерес в [8] представлява появата на ЛС при заетите лица, т.е. се включват единиците дали положителен отговор на въпроса: *„През МИНАЛАТА СЕДМИЦА работили ли сте някаква работа срещу заплащане или друг доход (поне 1 час)?”*. Друго важно разделение на единиците се прави чрез това дали заетостта е на пълно или непълно работно време. В анализа се включват само единици заети на пълно работно време и така признаците, обект на анализ, се редуцират до 26, а единиците регистрирали значения по тези признаци 48 529. Делът на ЛС при тези признаци

[1] Това съкращение е общоприето и използвано във всички езици за обозначаване на този алгоритъм, което подтиква автора да го използва и в българския език.

е нисък и въпреки това представлява особен интерес за наблюдение поради факта, че механизмът на тяхната поява е НСЛ. Признаците с по-значим дял на ЛС са: *Колко часа седмично работите ОБИКНОВЕНО на ОСНОВНАТА РАБОТА? (v14); Колко часа общо сте работили през МИНАЛАТА СЕДМИЦА на ОСНОВНАТА РАБОТА? (v22); Колко часа седмично желаете да работите - общо?(v25).*

В последните години EM алгоритъмът придобива все по-голяма популярност в статистическата обработка на информацията, независимо от предметната област на приложение. Наред със своята популярност той се „разпива" и намира свои деривати в други методи (например DA алгоритъмът на Tanner и Wang (1987)). За основатели на методът безспорно се посочват авторите Dempster, Lair и Rubin (1977) [1], въпреки че подобни идеи с десетилетия предхождат тази публикация. Meng и van Dyk показват в своята статия, че началните идеи на EM алгоритъмът могат да се открият в работата на McKendrick в далечната 1926 г., както и в идеите на Fisher 1925 г. [5].

Основната идея залегнала в EM алгоритъмът е да се даде решение на проблемът с липсващите стойности чрез последователно решаване на по-лесени проблеми базирани на пълната база данни.

EM алготитъмът е двустепенна итеративна процедура състояща се от E-стъпка и M-стъпка (E стъпката получила своето име от английската дума expectation, а M стъпката от думата maximization). Итеративната процедура започва с начални оценки на вектора на средните аритметични и ковариационната матрица. На тази начална итерация липсващите стойности се отстраняват постердвом премахване на единиците, при които са се появили. Така първоначално базата данни е от единици, за които има наблюдения при всички признаци пълна база данни. E стъпката използва елементите на вектора на средните и ковариационната матрица за да изгради система от регресионни уравнения, които да опишат променливите с ЛС. Целта на E - стъпката е да замести ЛС чрез регресионните зависимости между променливите. В последващата M-стъпка се прилагат последователно стандартните формули за получаване на средните аритметични и ковариационната матрица, но вече от пълната база данни от всички единици, включително и тези с попълнени значения от стъпка E. Алгоритъмът отново се връща към E-стъпката, използвайки получената информация от M-стъпката и наново оценя значенията на ЛС. Това се следва от M-стъпка, която обновява параметрите на базата данни средни и ковариации на базата на последните оценки за ЛС от E- стъпката и т.н.

Решаването на задачата посредством EM алгоритъма може да се опрости на изкачване на „върха" на функцията максималното правдоподобие [3]. Така началните оценки на средните и ковариациите в първата E стъпка могат да се вземат начални координати на изкачването. Последващите M и E стъпки са етапите на изкачването, като всяка следваща води до „издигане по-високо" по пътя към върха. При доближаване на пиковата точка стъпките стават все по малки и по-малки докато представат да са забележими и това се нарича конвергенция на модела. Когато промените във вектора на средните аритметични и ковариацията практически не настъпват промени през различните итерации максималната стойност на функцията на максималното правдоподобие е достигната и процесът приключва.

Представен в по-математизирана форма EM алгоритъмът изглежда по следния начин. За удобство се използва пример с две променливи X и Y, като ЛС се наблюдават само при Y. Познавайки механизмът на фукциониране на метода при двумерен модел той лесно може да се обобщи за многомерен.

Първоначалните условия за анализа се получават на базата на единиците, при които има наблюдения и за двете променливи. Тогава чрез

$$\hat{\mu}_Y = \frac{1}{N}\sum Y \qquad (1)$$

$$\hat{\sigma}_Y^2 = \frac{1}{N}\left(\sum Y - \frac{(\sum Y)^2}{N}\right) \qquad (2)$$

$$\hat{\sigma}_{XY} = \frac{1}{N}\left(\sum XY - \frac{\sum X \sum Y}{N}\right) \qquad (3)$$

се изчисляват максимално правдоподобните основни характеристики средна аритметична,

дисперсия и ковариация, необходими за оценката на регресионен модел със зависима величина Y и факторна променлива X. E-стъпката оценя значенията на Y чрез параметрите на модела:

$$\hat{\beta}_1 = \frac{\partial_{XY}}{\partial_X^2} \quad (4)$$

$$\hat{\beta}_0 = \hat{\mu}_Y - \hat{\beta}_1 \hat{\mu}_X \quad (5)$$

$$\hat{\sigma}_{Y|X}^2 = \partial_Y^2 - \hat{\beta}_1^2 \hat{\sigma}_X^2 \quad (6)$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (7)$$

От своя страна всяка E-стъпка е следвана от M стъпка, която използва новооценените значения на зависимата променлива Y и посредством формули 1 до 3 преоценя характеристиките на модела заложени във формули 4 до 7. Трябва да се има предвид, че E-стъпката не въвежда липсващите значения в базата данни, а само оценя очакваните стойности на Y в полза на изчисляването на основните характеристики на базата данни. Процесът приключва когато промените в основните характеристики станат клонящи към нула.

Алгоритъмът за увеличаване на данните, е много близък до извадките на Гибс (Gibbs sampler), който от своя страна е форма на Монте карло алгоритъма - MCMC (Markov chain Monte Carlo). Както е известно MCMC е сбор от техники за получаването на псевдо-случайни опити (избори) от вероятностни разпределения. През последните 30 години те са обект на по-особен интерес и са в основата на доста добри практически решения на много задачи. Основната цел на MCMC е генерирането на едно или повече значения на случайна променлива Z, която типично е многомерна [3:стр. 68]. Едно бързо описание на метода може да се направи по следния начин. Нека с P(Z) = f(Z) се означи плътността на Z, което се нарича и *целево разпределение на променливата.* Също така се генерира редицата $\{Z^{(1)}, Z^{(2)}, Z^{(3)},\ldots, Z^{(t)},\ldots\}$, в която всеки следващ елемент зависи от предходния по някакъв начин, и където граничното маргинално разпределение на $Z^{(t)}$, при t , е f. При t достатъчно голямо, $Z^{(t)}$ може да се разглежда като случайна извадка от f. Силата на MCMC е в случаите, когато е невъзможно или много трудно да се направят случайни опити директно от f, но е постижимо да се генерират редици от типа $Z^{(t)}$ за сяка изучавана променлива. Методът MCMC обикновено се класифицира като Бейсов или като „Бейсова постериорна симулация".

Извадките на Гибс (Gibbs sampler) са може би най-популярния вариант на MCMC. Методът е кръстен на американския физик, химик и математик Josiah Willard Gibbs (11.02.1839 г. 28.04. 1903 г.) от неговите автори братята Stuart and Donald Geman през 1984 г. [3], но получава своята популярност по-късно през 1990 чрез публикацията на Gelfand и Smith [4]. Същността на извадките на Гибс може да бъде предтавена накратко по следния начин.

Нека случайния вектор Z се раздели на J подвектора:

$$Z = (Z_1, Z_2, \ldots, Z_J).$$

Нека P(Z) представя разпределението на Z, което в определен анализа е необходимо да бъде симулирано. Посредством извадките на Гибс се правят последователни опити от зависимото разпределение на всеки подвектор при дадени всички останали. Тогава, при значение на Z вектора на стъпка t:

$$Z^{(t)} = \{Z_1^{(t)}, Z_2^{(t)}, \ldots, Z_J^{(t)}\},$$

може да се запише за стъпка (t+1):

$$Z^{(t+1)} = \{Z_1^{(t+1)}, Z_2^{(t+1)}, \ldots, Z_J^{(t+1)}\},$$

което се получава от последователните опити от условните разпределения:

$$\begin{aligned} Z_1^{(t+1)} &\sim P(Z_1 \mid Z_2^{(t)}, Z_3^{(t)}, \ldots, Z_J^{(t)}) \\ Z_2^{(t+1)} &\sim P(Z_2 \mid Z_1^{(t+1)}, Z_3^{(t)}, \ldots, Z_J^{(t)}) \\ &\cdot \\ &\cdot \\ &\cdot \\ Z_J^{(t+1)} &\sim P(Z_J \mid Z_1^{(t+1)}, Z_2^{(t+1)}, \ldots, Z_{J-1}^{(t+1)}) \end{aligned} \quad (8)$$

Казано по друг начин, на всяка итерация се правят извличане на значенията на Z_i от условните разпределения на Z_1, Z_2, ..., Z_J зависими всеки път от последните направени опити при останалитре условни разпределения на подвекторите.

След получаване на значенията на $Z^{(t+1)}$, по анлалогичен начин се продължава за $Z^{(t+2)}$,

$Z^{(t+3)}$, и т.н. Редицата $Z^{(t)}$ при t = 1, 2, ..., J образува марковска верига, която има разпределение идентично с P(Z) такова, че $Z(t) \rightarrow Z$, когато $t \rightarrow \infty$.

Алгоритъмът на увеличаване на данните е проява на извадките на Гибс представен от Tanner и Wang 1987 г. [7] специално за анализ на липсващи стойности. В увода на своята статия те описват модела като конкурентен на ЕМ алгоритъма за решаване на проблемът на максимизиране на функцията на правдоподобието, в случаите, когато тя не може да бъде апроксимирана чрез нормалното правдоподобие. В този случай максимално правдоподобните оценки и техните грешки не могат да се използват за получаването на валидни заключения. При представянето на метода трябва да се има предвид, че се разполага с случайния вектор z и той е разделен на два подвектора, z = (u, v), и разпределението P(z) не може да бъде симулирано, но условните разпределения P(u | v) = g(u | v) и P(v | u) = h(v | u) могат.

Следвайки логиката представена при извадките на Гибс, то ако се направи извадка с обем m от разпределението, което се апроксимира към целевото разпределиние - P(z), при итерация t може да се запише:

$$Z^{(t)} = (z_1^{(t)}, z_2^{(t)}, \ldots, z_m^{(t)}) = ((u_1^{(t)}, v_1^{(t)}), (u_2^{(t)}, v_2^{(t)}), \ldots, (u_m^{(t)}, v_m^{(t)}))$$

Тази извадка се възпроизвежда в две стъпки по отношение на u и v. Първата,

$$U^{(t+1)} = (u_1^{(t+1)}, u_2^{(t+1)}, \ldots, u_m^{(t+1)})$$

се получева чрез случаен независим подбор на $u_i^{(t+1)}$ от разпределението $g(u \mid v_i^{(t)})$, при i = 1, 2, …, m. Втората стъпка изисква случайна независима извадка при равна вероятност за подбор от композицията от зависимите разпределения $h(v \mid u_i^{(t+1)})$, от където се получава:

$$V^{(t+1)} = (v_1^{(t+1)}, v_2^{(t+1)}, \ldots, V_m^{(t+1)}) \text{ като}$$

$$\bar{h}(v \mid U(t+1)) = \frac{1}{m}\sum_1^m h(v \mid u_j^{(t+1)}), \qquad (9)$$

Така новата извадка при t+1 има вида:

$$Z^{(t+1)} = (z_1^{(t+1)}, z_2^{(t+1)}, \ldots, z_m^{(t+1)}) = ((u_1^{(t+1)}, v_1^{(t+1)}), (u_2^{(t+1)}, v_2^{(t+1)}), \ldots, (u_m^{(t+1)}, v_m^{(t+1)}))$$

Използвайки фукционален анализ Tanner и Wang доказват, че разпределението $Z^{(t)}$ конвергира към P(z), при $t \rightarrow \infty$. Размерът на m не е нужно да бъде голям. Като се вижда при m=1 DA алгоритъмът се трансформира във вариант на извадките на Гибс (8), при който J=2. Ако втората стъпка от алгоритъма се модифицира в независими опити от съответните условни разпрделения, вместо от композиция от разпределния, показан в (9), а именно:

$$v_i^{(t+1)} \sim h(v \mid u_i^{(t+1)}) \text{ за } i = 1, 2, \ldots, m$$

то алгоритъмът се превръща в m независими, паралелни приложения на извадките на Гибс [6, стр. 71]. Композицията от условните разпределения h(v | u) на всяка итерация могат да не дадат толкова добри практически резултати в скоростта на конвергенцията на $Z^{(t)}$, но когато m е голямо равенствата представени в (9) дават добра аналитична апроксимация на плътността , ако подобна апроксимация е желана.

Анализ на данните

Последователно се съпоставят резултати получени по трите матодики ЕМ алгоритъм, DA алгоритъм и DA алгоритъм при данни за началните параметри (средни и ковариационна матрица на променливите) от ЕМ алгоритъма (в изследването за краткост ще бъде означено с DA(em)). Както вече бе отбелязано, данните са разделени в две групи Кластър 7 и останали кластъри в зависимост от направената кластаризация на данните в Наблюдението на работната сила в България за 2007 г. в [8]. Основните характеристики средни аритметични и стандартни отклонения - на променливите при Кластър 7 са поместени в таб. 1. Те са оценени преди процедурата на въвеждане, но след оценка на параметрите при описаните анализи. Очевидна е близостта на резултатите при всички анализи.

Анализите се провеждат чрез NORM софтуерен пакет разработен от Schafer, J.L. за целите на обработката на ЛС. Това обяснява и някои различия в резултатите представени в това изследване и в [8], където изчисленията са направени чрез SPSS 16.

Първоначално се прилага **ЕМ алгоритъма**, който достига конвергенция при 62-рата итерация при оценка на максималното правдоподобие МП = -5704,37. Оценките на основните характеристики при променливите с

ЛС след направеното въвеждане от ЕМ аглоритъма са поместени в табл. 2. Прави впечатление снижаването на стандартната грешка в сравнение с първоначалните значения на променливите или т. н. наблюдавани стойности[2], както и екстремалните стойности в разпределенията по-големи от 90 часа на седмица. Диаграмата на разпределенията се показват в Граф. 1, а 95% доверителни интервали на средните аритметични на въведените стойности на Граф. 2. Изключително големия екцес на изходните данни е запазен и след въвеждането. Нещо, което не се вижда от представените резулатати, е че въведените стойности добре възпроизвеждат емпиричното разпределение на наблюдаваните стойности, но без да дават толкова голямо тегло на модалната стойност.

Таблица 1. Основни характеристики наблюдаваните стойности на променливите при приложените анализи за данните от Калстър 7

ПРОМЕНЛИВИ	СРЕДНИ АРИТМЕТИЧНИ И СТАНДАРТНИ ОТКЛОНЕНИЯ НА НАБЛЮДАВАНИТЕ ДАННИ		ПАРАМЕТРИ НА ПРОМЕЛИВИТЕ СЛЕД ЕМ		ПАРАМЕТРИ НА ПРОМЕНЛИВИТЕ СЛЕД DA(ем) - 1000 ИТЕРАЦИИ		ПАРАМЕТРИ НА ПРОМЕНЛИВИТЕ СЛЕД DA - 1000 ИТЕРАЦИИ	
	СР. АРИТ.	СТ. ОТКЛ.	СР. АРИТ.	СТ. ОТКЛ.	СР. АРИТ.	СТ. ОТКЛ.	СР. АРИТ.	СТ. ОТКЛ.
v14	**61.3580**	**7.72655**	**61.3512**	**7.60949**	**61.2138**	**7.42120**	**61.2138**	**7.42120**
v15	1.97636	0.157626	1.97636	0.157591	1.98102	0.160776	1.98102	0.160776
v16	0.255692	1.66417	0.255692	1.66380	0.209866	1.70139	0.209866	1.70139
v17	1.91900	0.355185	1.91900	0.355107	1.92362	0.364516	1.92362	0.364516
v18	0.0547285	0.334967	0.0547285	0.334893	0.0641897	0.335863	0.0641897	0.335863
v19	1.97767	0.204913	1.97767	0.204868	1.97627	0.206314	1.97627	0.206314
v20	0.156305	1.83856	0.156305	1.83815	0.153062	1.83004	0.153062	1.83004
v21	0.128284	1.40449	0.128284	1.40418	0.122177	1.39399	0.122177	1.39399
v22	**61.4314**	**8.05913**	**61.3635**	**8.15735**	**61.3244**	**8.15384**	**61.3244**	**8.15384**
v23	2.93564	0.281989	2.93564	0.281927	2.93771	0.283379	2.93771	0.283379
v24	0.0223292	0.251026	0.0223292	0.250971	0.0215159	0.254036	0.0215159	0.254036
v25	**59.9337**	**8.92692**	**60.3989**	**8.30762**	**60.3729**	**8.35125**	**60.3729**	**8.35125**
v26bgr	5.85902	3.21684	5.85902	3.21613	5.73158	3.19571	5.73158	3.19571
v26va	2002.06	4.92017	2002.06	4.91909	2001.85	4.96252	2001.85	4.96252
v26vb	2.05035	3.38060	2.05035	3.37986	1.98264	3.38602	1.98264	3.38602
v26g	0.699212	0.938391	0.699212	0.938185	0.681140	0.947798	0.681140	0.947798
v26d2	1.09238	0.289628	1.09238	0.289564	1.08626	0.301350	1.08626	0.301350
v26d3	0.115149	0.385181	0.115149	0.385097	0.105337	0.402799	0.105337	0.402799
v26d4	0.789405	3.56859	0.789405	3.56781	0.723341	3.72948	0.723341	3.72948
v26egr	0.0245184	1.17176	0.0245184	1.17151	0.0307248	1.18481	0.0307248	1.18481
v26j	5.42338	2.59768	5.42338	2.59711	5.53314	2.62544	5.53314	2.62544
v26z	1.76795	0.423269	1.76795	0.423176	1.77791	0.423184	1.77791	0.423184
v26k	1.84151	1.00298	1.84151	1.00276	1.81641	1.01311	1.81641	1.01311

[2] Тази част от променливите, за които има регистрирана информация при Наблюдението на работната сила. Наблюдаваните стойности заедно с въведените ЛС правят базата данни пълна.