Журнал Американской статистической ассоциации

ISSN: 0162-1459 (печать) 1537-274X (онлайн) Домашняя страница журнала: http://www.tandfonline.com/loi/uasa20

Устойчивый неправильный максимум правдоподобия: настройка, Вычисления и сравнение с другими Методы робастной гауссовской кластеризации

Пьетро Коретто и Кристиан Хенниг

Страница 2

ЖУРНАЛ АМЕРИКАНСКОЙ СТАТИСТИЧЕСКОЙ АССОЦИАЦИИ
□□□□, ТОМ. □□□, НЕТ. □□□, □□□□ – □□□□, Теория и методы
http://dx.doi.org/□□.□□□□/□□□□□□□□.□□□□.□□□□□□□

Устойчивый неправильный максимум правдоподобия: настройка, вычисление и сравнение с Другие методы робастной гауссовской кластеризации

Пьетро Коретто [a] и [redacted]истиан Хенниг [б]

Департамент экономики и статистики, Университет Салерно, Fisciano, Италия; [b] Департамент статистических наук Университетского колледжа Лондона, Лондон, Великобритания

АБСТРАКТНЫЙ

Две основные темы этой статьи - введение «оптимально настроенного надежного несобственного максимума. оценка правдоподобия »(OTRIMLE) для надежной кластеризации на основе многомерной гауссовской модели для кластеров, и всестороннее исследование моделирования, сравнивающее OTRIMLE с максимальной вероятностью в гауссовой смеси. диаграммы с шумовой составляющей и без нее , смеси *t*- распределений и подход TCLUST для усеченных кластеризация. OTRIMLE использует неправильную постоянную плотность для моделирования выбросов и шума. Это может быть выбирается оптимально, чтобы часть данных, не содержащая шумов, выглядела как можно ближе к гауссовой смеси. Некоторое отклонение от гауссовости можно обменять на снижение расчетной доли шума. Ковариация также рассматриваются матричные ограничения и вычисление OTRIMLE. При моделировании все методы сталкиваются с установками, в которых их предположения модели не выполняются в точности, и для оценки экспериментов стандартизированным способом путем неправильной классификации, новое определение, основанное на моделях, «истинных кластеров». ters », что отклоняется от обычного отождествления компонентов смеси с кластерами. в исследования, каждый метод оказывается лучше для одной или нескольких установок, но OTRIMLE достигает наибольшего удовлетворительная работа работы. Эти методы также применяются к двум реальным наборам данных, одному без и другому. с известными «истинными» кластерами. Дополнительные материалы к этой статье доступны в Интернете.

СТАТЬЯ ИСТОРИЯ
Поступила в июле □□□□
Пересмотрено в сентябре □□□□

КЛЮЧЕВЫЕ СЛОВА
Кластерный анализ; EM-алгоритм; Неправильный плотность; Модели смесей; Кластеризация на основе моделей; Надежность

1. Введение

В этой статье мы представляем и исследуем «оптимально настроенные надежная неправильная оценка максимального правдоподобия »(OTRIMLE), метод устойчивой кластеризации с кластерами, которые могут быть аппроксимируется многомерными распределениями Гаусса. Его одинразмерная версия была представлена Коретто и Хенниг (2010 г.). Мы также представляем имитационное исследование, сравнивающее OTRIMLE и другие подходы для (в основном надежной) модели - на основе кластеризации, которая, насколько нам известно, является наиболее сложной тщательное изучение в этой области и включает в себя тщательное обсуждение проблема сравнения методов, основанных на различных модельных предположениях.

Основная идея OTRIMLE - подогнать неподходящую плотность к данные, состоящие из гауссовой плотности смеси и «Компонент псевдосмеси», определяемый небольшой постоянной плотностью sity, который предназначен для улавливания выбросов в областях с низкой плотностью данные. Это навеяно добавлением однородного «шума компонент »к гауссовской смеси (Banfield and Raftery 1993). Хенниг (2004 г.) показал, что использование неподходящей плотности улучшает отказоустойчивость этого подхода. OTRIMLE имеет было обнаружено, что он хорошо работает с одномерными данными в Coretto и Хенниг (2010 г.).

Как и во многих других статистических задачах, нарушения модели предположения и особенно выбросы могут вызвать проблемы в кластерный анализ. Наше общее отношение к использованию статистических моделей в кластерном анализе заключается в том, что модели не должны быть понимается как отражение некоторого основного, но на практике необслужащая «истина», а скорее как мысленные конструкции, подразумевающие достоверность поведение методов, производных от них (например, максимизация вероятность), который может или не может быть уместным в данном приложение (подробнее об общей философии кластеринг можно найти в Hennig and Liao 2013). Используя такую модель как смесь многомерных гауссовских распределений, интерпретируйтерассмотрение каждого компонента смеси как «кластера» означает, что мы смотрим для кластеров, которые имеют приблизительно «гауссову форму», но мы не хотят полагаться на то, действительно ли данные были сгенерированы по гауссовой смесью. Мы заинтересованы в спектакле таких методов в ситуациях, когда можно законно искать для кластеров гауссовой формы, даже если некоторые точки данных не принадлежат таким кластерам (далее именуемым «шумом»), и даже если кластеры не совсем гауссовы. Это отражает факт что на практике, например, используются смеси *t* -распределений для кластеризации тех же наборов данных, к которым относятся гауссовские смеси подобраны, интерпретируя полученные кластеры таким же образом.

Для иллюстрации проблемы выбросов в кластерах на основе моделей: Мы используем пятимерный набор данных, в котором 170 отсутствуют. районы немецкого города Дортмунд характеризуются количество переменных, которое подробно обсуждается в разделе 6.1. Подгонка простой гауссовой смеси с $G = 4$ ко всем пяти переменным от пакета R MCLUST (Fraley et al. 2012), один кластер является одноточечный кластер, состоящий только из крайнего выброса и двух дополнительных эти кластеры соответствуют двум различным разновидностям умеренных выбросов. Более

СВЯЗАТЬСЯ с Пьетро Коретто pcoretto@unisa.it Департамент экономики и статистики, Университет Салерно, □□□□□ Fisciano SA, Италия.
Цветные версии одного или нескольких рисунков в статье можно найти в Интернете по адресу *www.tandfonline.com/r/JASA*.
Дополнительные материалы к этой статье доступны в Интернете. Пожалуйста , перейдите на www.tandfonline.com/r/JASA.

Стр. 3

в единый кластер собрано более 120 районов. Задача надежная кластеризация позволяет избежать множества или даже большинства кластеров преобладают выбросы, и для создания значимой кластеризации структура также среди основной массы неэкстремальных наблюдений.

Ряд методов кластеризации на основе моделей, которые могут с выбросами были предложены в последние годы. Обзор из этих методов приведено в разделе 2.. OTRIMLE - это введение дается и обсуждается в разделе 3., начиная с «RIMLE», в котором уровень несобственной постоянной плотности является настройкой постоянный. Затем мы представляем метод оптимальной настройки и обсудить его вычисление. В разделе 5 представлены результаты моделирования, который использует единый подход для определения кластеров эллиптической формыздесь представлены все методы кластеризации на основе моделей. gmix это реализовано в пакете R MCLUST (Fraley et al. 2012). В виде проиллюстрировано в Разделе 1 и доказано в Хенниге (2004 г.), метод может сильно зависеть от выбросов и отклонений от предположения модели, и теперь мы переходим к подходам, которые пытаются разобраться с этой проблемой. Из-за нехватки места мы представляем их в подробно в онлайн-приложении и дайте только краткий обзор здесь.

2.2. Оценщик типа ML для гауссовских смесей с Равномерный шум (gmix.u)

индикаторы с шумом / выбросами, представленные в разделе 4. Дортмунд упомянутый выше набор данных обсуждается в разделе 6 вместе с набор мелодий народных песен из двух известных регионов. Некоторый дальнейшие вопросы, включая оценку количества кластеров обсуждаются в разделе 7. Дополнительные сведения об экзамене-набор данных (включая полные диаграммы рассеяния), исследование моделирования, и расчет методов представлены в онлайн-приложении мент (Коретто и Хенниг в прессе b). Теоретические свойства RIMLE с фиксированной постоянной настройки исследуются в Коретто и Хенниг (в прессе) и цитируется здесь.

2. Методы из литературы.

Далее предположим, что наблюдаемый образец $x_n = \{x_1, x_2, ..., x_n\}$, где x_i - реализация случайной величины $X_i \in \mathbb{R}^p$ с $p \geq 1$; $X_1, ..., X_n$ iid. Цель состоит в том, чтобы сгруппировать точки выборки в G различных групп.

2.1. Максимальное правдоподобие (ML) для гауссовских смесей (gmix)

Пусть $\varphi(x; \mu, \cdot)$ - плотность многомерного гауссовского распределения. вычисление со средним вектором $\mu \in \mathbb{R}^p$ и ковариационной матрицей $p \times p$. Предположим, что наблюдаемый образец взят из конечной Распределение гауссовой смеси с плотностью

$$m(x; \theta) = \sum_{j=1}^{грамм} \pi_j \varphi(x; \mu_j, {}_j), \quad (2.1)$$

где $\pi_j \in [0, 1]$ для всех $j = 1, 2, ..., G$ и $\sum_{j=1}^{G} \pi_j = 1$, θ - это вектор параметров, содержащий тройки $\pi_j, \mu_j, {}_j$ для всех $J = 1, 2, ..., G$. Кластеризация совпадает с присвоением точек компоненты смеси на основе оценок параметров ML. π_j можно интерпретировать как ожидаемую долю начисленных баллов. получено из j- го компонента. Пусть θ^{ml}_n быть оценкой ML-тор для θ, обычно вычисляемый путем максимизации математического ожидания (EM) алгоритм (Dempster et al. 1977). Оценщик ML под (2.1) существует только при соответствующих ограничениях на ковари-матрицы ances. Эти ограничения (которые также актуальны для методы, представленные ниже) будут подробно обсуждены в Раздел 3. Пусть τ^{ml}_{ij} - оценочная апостериорная вероятность того, что наблюдаемая точка x_i была получена из j- го компонента смеси. nent, то есть

$$\tau^{\text{мл}}_{ij} = \frac{\pi^{\text{мл}}_{j,n} \varphi(x_i; \mu^{\text{мл}}_{j,n}, {}^{\text{мл}}_{j,n})}{m(x_i; \theta^{\text{ml}}_n)} \quad \text{для всех} \quad j = 1, 2, ..., G. \; (2.2)$$

Затем точка x_i может быть отнесена к k- му кластеру, если $k = \arg\max_{j=1,2,...,G} \tau^{\text{ml}}_{ij}$. Этот метод присваивания является общим для

Банфилд и Рафтери (1993 г.) добавили однородную смесь компо-на наименьшем гипер прямоугольнике, покрывающем данные до (2.1), называя это « *шумовой составляющей* », чтобы приспособиться к «шуму».

2.3. ML для смесей t-распределений Стьюдента (tmix)

Маклахлан и Пил (2000 г.) заменил гауссовские плотности в (2.1) С многомерным Студентом - m плотностью, потому что они имеют более тяжелые хвосты и, следовательно, могут вместить выбросы в лучшую способ. Наблюдения можно назвать «шумовыми», если они проводятся в невысокой берлоге. Область t- распределения соответствует их кластеру. Хенниг (2004 г.) показал, что ни tmix, ни gmix.u не являются отказоустойчивыми.

2.4. TCLUST

TCLUST основан на максимизации усеченной вероятности «модель фиксированного разбиения» с весами кластеров π_j. При $R = \cup_{j=1}^{G} R_j$, $\#\{R\} = [n(1-\alpha)]$ количество необрезанных точек:

$$\theta^{\text{tclust}} \; \text{знак равно} \; \arg\max_{\theta \in, \#\{R\} = [n(1-\alpha)]} \sum_{j=1}^{грамм} \sum_{i \in R_j} \left(\text{журнал} \; \pi_j + \text{журнал} \; \varphi(x; \mu_j, {}_j) \right)$$

(2.3)

Для получения дополнительной информации см. Гальегос (2002 г.), Гальегос и Риттер (2005), и Гарсия-Эскудеро и др. (2008 г.). Метод TCLUST-ogy реализован в пакете TCLUST R от Fritz, García-Эскудеро и Майо-Искар (2012 г.). Методы разбиения с помощью обрезка началась с предложения обрезанных k- средних Куэста-Альбертос, Гордализа и Матран (1997 г.).

2.5. Дальнейшие существующие работы

Можно найти больше подходов к надежной кластеризации на основе моделей в литературе. Нейков и др. (2007 г.) предложено и реализовано метод усеченного правдоподобия. Цинь и Прибе (2013) вступление-разработал EM-алгоритм, адаптированный к максимальной L_q- вероятности оценка и изучила его поведение в рамках модели грубых ошибок. Ссылки на другие подходы к устойчивой кластеризации приведены в Гарсия-Эскудеро и др. (2010 г.).

3. Оптимально настроенный устойчивый неправильный максимум Вероятность

3.1. Устойчивый неправильный максимум правдоподобия

Робастная неправильная оценка максимального правдоподобия (RIMLE) основан на идее «шумовой составляющей» робастизации.

Стр. 4

(gmix.u). Основная идея - использовать псевдомодель, в которой шум представлен неправильной постоянной плотностью по все евклидово пространство:

$$\psi_\delta(x, \theta) = \pi_0 \delta + \sum_{j=1}^{грамм} \pi_j \varphi(x; \mu_j, {}_j), \quad (3.1)$$

с $\pi_0, \pi_j \in [0, 1]$ для $j = 1, 2, ..., G$, $\pi_0 + \sum_{я=1}^{G} \pi_j = 1$. $\delta > 0$ - несобственная постоянная плотность (icd). В вектор параметров θ содержит все гауссовские параметры плюс все параметры пропорции, включая π_0, т. е. $\theta = (\pi_0, \pi_1, ..., \pi_G, \mu_1, ..., \mu_G, {}_1, ..., {}_G)$. Учитывая сам-простая неправильная функция псевдоблогарифмического правдоподобия

$$l_n(\theta) = \frac{1}{n} \sum_{я=1}^{n} \text{журнал} \; \psi_\delta(x_i, \theta), \quad (3,2)$$

RIMLE определяется как

$$\theta_n(\delta) = \arg\max_{\theta \in} \; l_n(\theta), \quad (3.3)$$

где является подходящим пространством с ограниченными параметрами dis-ругали ниже. $\theta_n(\delta)$ затем используется для кластеризации точек, как для модели - методы кластеризации на основе. Определите псевдо-апостериорные вероятности по аналогии с (2.2):

$$\tau_j(x_i, \theta) := \begin{cases} \dfrac{\pi_0 \delta}{\psi_\delta(x_i, \theta)} & \text{если } j = 0 \\ \dfrac{\pi_j \varphi(x_i, \mu_j, {}_j)}{\psi_\delta(x_i, \theta)} & \end{cases} \quad \text{для } i = 1, 2, ..., n,$$

(1981 г.) и изучен Хэтэуэем (1985).) для одномерных Гауссовы смеси. EM-алгоритмы вычисления ML многомерные гауссовские смеси при (3.5) были изучены автор: Ingrassia (2004 г.) и Инграссия и Роччи (2007 г.), хотя асимптотические свойства соответствующего МЛЭ не были изучены. доказано. Те же ограничения используются для TCLUST Гарсиаа. Escudero et al. (2008 г.). Есть несколько альтернативных вариантов. напряжения, см. Ingrassia и Rocci (2011 г.) и Гальегос и Риттер (2009 г.).

Хотя (3.5) предотвращает несвязанность вероятности в моделях стандартной смеси и TCLUST, для RIMLE это не достаточно. Точки, не подгоняемые ни одной из гауссовских компонент, могут все еще может быть подогнан неподходящим однородным компонентом. Следовательно, Коретто и Хенниг (в прессе) предложил дополнительный «шум ограничение пропорции, "

$$\frac{1}{n} \sum_{я=1}^{n} \tau_0(x_i, \theta) \leq \pi_{\max}, \quad (3,6)$$

при фиксированном $0 < \pi_{\max} < 1$. Величина $n^{-1} \sum_{i=1}^{n} \tau_0(x_i, \theta)$ может быть интерпретируется как расчетная доля точек шума. Параметр $\pi_{\max} = 0.5$ просто реализует знакомое условие в надежной статистике. тиков, что не более половины данных следует классифицировать как «выходящие из строя». лжи / шум ». Результирующее ограниченное пространство параметров для RIMLE затем

$$\text{знак равно} \left\{ \theta : \pi_j \geq 0 \; j \geq 1; \; \pi_0 + \sum_{j}^{грамм} \pi_j = 1; \; \frac{1}{} \sum^{n} \tau_0(x_i, \theta) \right.$$

$\psi_\delta(x_i, \theta)$ если $j = 1, 2, ..., G$;

и назначаем баллы согласно следующему правилу

$$J(x_i, \theta) := \arg\max_{j \in \{0, 1, 2, ..., G\}} \tau_j(x_i, \theta). \quad (3,4)$$

Зафиксировав δ, (3.1) не определяет правильную вероятностную модель, но (3.3) дает полезную процедуру для данных, смоделированных как пропорции ции $(1 - \pi_0)$ смеси гауссовых распределений плюс доля π_0 баллов, не отнесенных ни к одному значимому кластеру. Области высокой плотности скорее связаны с кластерами, чем с с шумом, поэтому области шума должны быть с наименьшими плотность. Этого можно добиться, используя однородную плотность как в gmix.u, но для этого наличия сильных выбросов зависимость в силу равномерного распределения на выпуклой оболочке данных по-прежнему вызывает проблему устойчивости (Hennig 2004).

Задача оптимизации в (3.3) требует, чтобы подходит определено, иначе $\theta_n(\delta)$ может не существовать. Как обнаружил Дэй (1969 г.), вероятность гауссовой смеси может вырождаться. Этот проблема распространяется на (3.1) также. Пусть $\lambda_{k,j}$ - собственное значение j для некоторого $K = 1, 2, ..., p$ и $J = 1, 2, ..., G$. Возьмите последовательность $(\theta_m)_{m \in N}$ такое, что $\lambda_{k,j,m} \searrow 0$ и $\mu_{j,m} = x_1$, то $l_n(\theta_m) \to + \infty$. Есть несколько способов избежать этой проблемы. Пусть $\lambda_{\max}(\theta)$ и $\lambda_{\min}(\theta)$ - соответственно максимальное и минимальное собственное значения ковариационных матриц в θ, Коретто и Хенниг (в нажмите а) приняла «ограничение на собственное отношение»

$$\lambda_{\text{макс}}(\theta) / \lambda_{\text{мин}}(\theta) \leq \gamma < +\infty \quad (3.5)$$

с фиксированным $\gamma \geq 1$. $\gamma = 1$ ограничивает все компоненты ковариации матрицы должны быть сферическими и равными, как в кластеризации k- средних а $\gamma > 1$ ограничивает относительное расхождение кластеры. Этот тип ограничения был предложен Деннисом.

$$\leq \pi_{\max}; \quad \frac{\lambda_{\text{макс}}(\theta)}{\lambda_{\text{мин}}(\theta)} \leq \gamma \quad . \quad (3,7)$$

Коретто и Хенниг (в прессе) показал, что $\theta_n(\delta)$ существует для любого $\delta \geq 0$, если $\#(x_n) > G + \lceil n\pi_{\max} \rceil$ и что $\theta_n(0)$ существует при более мягкое условие , что $\#(x_n) > G$. При $\delta = 0$ RIMLE сводит до ML для простых гауссовских смесей. Пусть $E_P f(x)$ - математическое ожидание. ции $F(x)$ при $x \sim P$. Функционал RIMLE определяется как

$$\theta_*(\delta) = \arg\max_{\theta \in G} E_P \text{ журнал } \psi_\delta(x; \theta). \quad (3.8)$$

Существование (3.8), согласованность $\theta_n(\delta)$ на факторпространстве топология, идентифицирующая все максимумы логарифмического правдоподобия и ее нарушение точки вниз показаны у Коретто и Хеннига (в прессе).

3.2. Оптимальный неправильный уровень плотности

Иногда знание предмета может быть доступно для помощи: при выборе δ, но такие ситуации довольно редки. Здесь мы предлагаем выбор δ *в* зависимости от данных . Обратите внимание, что δ не рассматривается как модельная величина, подлежащая оценке здесь, а скорее как настраивающее устройство для обеспечения надежной кластеризации. Цель RIMLE - это аппроксимация плотности кластерных областей точки, когда эти области выглядят как области, созданные Gaus- сианское распределение. Определим «оптимальное» значение δ как минимизатор целевой функции, измеряющей невязку найденных кластеры из гауссовского прототипа. Для данного $\theta_n(\delta)$ определим кластерный квадрат расстояния Махаланобиса до центров кластеров в виде

$$d_{i,j,n} = (x_i - \mu_{j,n})' \Sigma^{-1}_{j,n} (x_i - \mu_{j,n}),$$

Стр. 5

и кластерное взвешенное эмпирическое распределение $d_{i,j,n}$,

$$M_j(t; \delta) = \frac{1}{\sum_{i=1}^n \tau_j(x_i, \theta_n(\delta))} \sum_{я=1}^n \tau_j(x_i, \theta_n(\delta)) \mathbf{1}\{d_{i,j,n} \leq t\},$$
$$j = 1, 2, ..., G. \quad (3.9)$$

В M_j *расстояние* до i- й точки взвешивается в соответствии с псевдо-апостериорная вероятность того, что i- е наблюдение было генерируется j- м компонентом смеси. Если j- й кластер приблизительно по Гауссу и $\mu_{j,n}$ и $\Sigma_{j,n}$ хороши приблизительно изображения его местоположения и разброса, мы ожидаем, что в квадрате Расстояния Махаланобиса до $\mu_{j,n}$ точек, действительно принадлежащих компонент смеси № j (для которого $\tau_j(\cdot)$ указывает оценку сопряженная вероятность) будет приближаться к χ^2_p распределение. С участием $\chi^2_p(a)$ - значение cdf χ^2_p распределение в a, определим расстояние типа Колмогорова для j- го кластера

$$K_j(\delta) = \text{макс.}_{я = 1, ..., n} \left| M_j(d_{i,j,n}; \delta) - \chi^2_n(\partial_{я,j,n}) \right|. \quad (3.10)$$

Затем оценивается качество общего гауссовского приближения: рассчитывается путем взвешивания $K_j(\cdot)$ с расчетным соотношением компонентов Тион $\pi_{J,n}$:

$$D(\delta) = \frac{1}{\sum_{j=1}^G \pi_{j,n}} \sum_{j=1}^{\text{грамм}} \pi_{j,n} K_j(\delta). \quad (3.11)$$

Для данной константы $\beta \geq 0$ определите оптимальный уровень icd как

$$\delta_n = \arg\min_{\delta \in [0, \delta_{\max}]} \left[D(\delta) + \beta\pi_{0,n} \right]. \quad (3.12)$$

Соответствующая оптимально настроенная RIMLE (OTRIMLE) будет обозначается как $\theta_n(\delta_n)$. Существование и единственность δ_n не тривиальны ial, см. раздел 3.3 . Раздел 5.4 посвящен тому, как $D(\delta)$ ведет себя как функция δ.

Простой выбор β, формализующий «гауссовский кластер », $\beta = 0$. Однако на практике часто это не так. настолько важно, чтобы кластеры имели точную гауссову форму насколько возможно. $\beta > 0$ (но обычно меньше 1) формализует, что менее гауссовские формы кластеров допускаются, если это приводит к расчетная доля шума снижена. Как видно в разделе 5, выбирая $\beta = \frac{1}{3}$ приводит к улучшениям, если истинные кластеры t -распределенный. Раздел 5.5 дает более подробную информацию о влиянии

Стропила (1993 г.) затем используется для нахождения начальных гауссовских кластеров среди шума. См. Онлайн-приложение для случая, когда найденный раздел недействителен.

OTRIMLE можно найти, вычислив RIMLE на сетка значений δ от нуля до некоторого достаточно большого $\delta_{\max}$. На практике решаем программу (3,12) по «золотому сечению» поиск »Кифера (1953 г.) по множеству кандидатов $\delta \in [0, \delta_{\max}]$. В В большинстве численных экспериментов мы обнаружили, что не более 30 Требуются оценки RIMLE. $\delta_{\max}$ можно выбрать как наивысшее значение плотности, возникающее в инициализированном кластере, отбрасывая δ -значения, при которых RIMLE-решение заканчивается на границе пространство параметров.

4. Определение «истинных» кластеров и неправильная классификация

В большинстве имитационных исследований в кластерном анализе данные генерируются из смешанных (или фиксированных) моделей, это предполагается, что «истинные» кластеры отождествляются со смесью компоненты и методы могут быть затем оценены ошибочной классификацией. коэффициенты классификации. Но это может быть проблематично. Рассмотрим сравнение оценок ML для гауссовых смесей и для смеси t -распределений. В большинстве приложений оба подходы будут рассматриваться как потенциально подходящие для того же, а именно для поиска кластеров, которые являются единообразными. модальные и эллиптические. В приложениях с кластеризацией основной интерес (в отличие от оценки параметров) исследователей было бы неважно, будет ли плотность вокруг их скоплений- Термические ядра скорее выглядят как гауссово или t -распределение. Но последствия, для которых точки считаются «истинными выбросами» по сравнению с «действительно принадлежащим кластеру» будет другим, потому что некоторые точки, порожденные t -распределением с низкими степенями свобода действительно выходит за рамки ядра t -распределение, из которого они генерируются.

В более общем плане идентификация кластеров и смесей компоненты нельзя воспринимать как должное. Хенниг (2010 г.) иллю- Было обнаружено, что интерпретация компонентов гауссовой смеси как кластеры, зависит от того, разделены ли компоненты достаточно. А в устойчивой кластеризации часто можно интерпретировать группу из нескольких точек с низкой плотностью как «шум», даже если они были сгенерированы распределением Гаусса.

Теперь мы определим «эталонную истину» для моделей смеси. которые используются в нашем исследовании моделирования, из которых ошибочно классифицируются выбирая $\beta > 0$.

3.3. Вычисление

Для фиксированного δ RIMLE можно соответствующим образом вычислить, используя алгоритм ожидания-максимизации (EM). См. Коретто и Хенниг (в прессе) для получения подробной информации и онлайн-приложение для подробности и предыстория следующего. Итог EM-алгоритм зависит от инициализации. Мы использовали начальный Метод tialization, вдохновленный программным обеспечением MCLUST. Избегать ложные кластеры, мы считаем действительными только начальные разделы те, которые содержат не менее min.pr × n наблюдений в каждом кластере - тер (min.pr = 0 . 005, скажем). В качестве первой попытки найти такой действующий разделение, обнаружение помех / шума на основе ближайшего соседа поставлено Байерсом и Рэфтери (1998 г.) применяется для идентификации начального шумовая догадка. Агломеративная иерархическая кластеризация на основе максимального обучения критерии для моделей гауссовой смеси, предложенные Банфилдом и

Затем можно рассчитать скорость фиксации. Для мотивации рассмотрим фигура 1, который показывает искусственный набор данных, взятый из смеси двух гауссианов и равномерного распределения. фигура 1а) показывает немаркированный набор данных, с которым работает кластерный анализ. фигура 1(б) показаны точки, помеченные составом смеси. Ненты, которые их породили. Обратите внимание на три красные звезды (генерируется из равномерного шума) в середине области где два гауссианца имеют большую часть своей массы. Более того, есть зеленые точки от левой гауссовой компоненты с более низкая плотность, которая попадает в плотную синюю область. Нет метода можно ожидать реконструировать все членство в кластерах в такие перекрывающиеся регионы. фигура 1(с) показывает то, что мы определяем как «справочную истину», определено далее.

Идея состоит в том, что мы выбираем вероятностные меры $P_1, P_2, \dots, P_G$, чтобы соответствовать G «истинным кластерам» (подразумевая что они «группируются», то есть генерируют четко различимые,

Стр. 6

Рисунок . Искусственный набор данных, состоящий из точек, взятых из двух двумерных распределений Гаусса, и точек из равномерного распределения на квадрате. [-10 , 10] × [-5 , 15]. (a) немаркированные точки, (b) окрашенные в соответствии с компонентами смеси, (c) цвета представляют два GR $_\alpha$ (при $\alpha = 10^{-4}$); красные звезды принадлежат NR.

хотя и не обязательно неперекрывающиеся, шаблоны данных). Для каждый из них, мы определяем область точек, которые могут быть рассмотрены определяется как не выпадающие значения на основе среднего и ковариационной матрицы функционал, согласованный по Фишеру на гауссовском неравенстве tribution, но надежный и существующий и для других дистрибутивов. Это определяет область не выбросов гауссовой формы. Мы считать все точки «шумом», которые в соответствии с этим определение для всех $P_1, P_2, \dots, P_G$. Квадратичный дискриминантный анализ присваивает точки кластерам, которые не являются выбросами по отношению к более чем один из P_j. Это означает, что баллы присваиваются кластеры по оптимальным границам классификации по гауссову предположение, даже если компоненты на самом деле не гауссовы. Это формализует использование «прототипов гауссовских кластеров» без предполагая, что кластеры действительно должны быть гауссовскими.

Для этого пусть m_j и S_j будут минимальным определителем ковариации. минант (MCD) центр и функционал рассеяния при P_j (Rousseeuw 1985 г.). Катор и Лопухаа (2012 г.) доказал существование МКД функционал для широкого класса вероятностных мер. S_J может быть исправлено для достижения согласованности при P_j, равном гауссову распределение (Croux and Haesbroeck 1999; Писон и др. 2002 г.), так что, когда P_j является гауссовским, m_j и S_j являются соответствующими вектор среднего и ковариационная матрица. Пусть π_j - ожидаемая про- часть точек, порожденных P_j. Мы разрешаем $\sum_{j=1}^{G} \pi_j \leq 1$, так что точки могут быть сгенерированы (шума) распределениями других чем $P_1, P_2, \dots, P_G$. Определите квадратичный дискриминантный балл для назначая точку $y \in \mathbb{R}^p$ к J - й кластер за счет максимального

$$\operatorname{qs}(y; \pi_j, m_j, S_j) := \text{журнал}(\pi_j) - \frac{1}{2}\text{журнал}(\det(S_j)) - \frac{1}{2}(y - m_j)' S_j^{-1}(y - m_j),$$

для $j = 1, 2, \dots, G$.

Если кластеры действительно гауссовы, это эквивалентно (3,4). Рассмотреть возможность

$$c_\alpha(m_j, S_j) := \{ y : (y - m_j)' S_j^{-1}(y - m_j) \leq \chi^2_p(1 - \alpha) \},$$

где χ^2 из $\chi^2_p(1-\alpha)$, поскольку для гауссовского P_j, $P_j(\mathbb{R}^p \setminus c_\alpha(m_j, S_j)) = \alpha$. Для фиксированного уровня α область α -Гаусса определяется как объединение этих эллипсоидов:

$$\mathrm{GR}_\alpha := \bigcup_{j=1}^{\text{грамм}} c_\alpha(m_j, S_j),$$

а область шума определяется выражением $\mathrm{NR}_\alpha := \mathbb{R}^p \setminus \mathrm{GR}_\alpha$.

Определение 1 (принадлежность к α-гауссовскому кластеру) . Для данного $\alpha \in [0, 1)$, процесс генерации данных (DGP) с параметром кластера ters $\theta_C := \{ (\pi_j, m_j, S_j), j = 1, 2, \dots, G \}$, и набор данных $x_n := \{ x_1, x_2, \dots, x_n \}$; дается членство в α- гауссовском кластере от

$$\mathrm{AGR}_\alpha(x_i; \theta_C) := \mathbf{1}\{ x_i \in \mathrm{GR}_\alpha \} \times \arg\max_{j=1,2,\dots,G} j\, \mathbf{1}\{ j = \arg\max_{g=1,\dots,G} \operatorname{qs}(y; \pi_g, m_g, S_g) \}. \tag{4.1}$$

$\mathrm{AGR}_\alpha(x_i, \theta_C) = 0$ означает, что $x_i \in \mathrm{NR}_\alpha$.

Это определение основано на определении выбросов с относительно эталонной модели, как у Дэвиса и Гэзера (1993 г.) и Becker and Gather (1999 г.). Разница в том, что здесь параметр α не контролирует напрямую вероятность область шума. Если α и тройки (π_j, m_j, S_j) зафиксированы для все $j = 1, 2, \dots, G$, размер области шума будет зависеть от степени перекрытия и гауссовости эллипсоидов в GR_α. α должен быть малым, потому что идея выброса подразумевает что при распределении Гаусса выбросы очень редки. Мы выберите $\alpha = 10^{-4}$, откуда следует, что вероятность того, что существует по крайней мере один выброс в $n = 500$ гауссовских наблюдениях 0 . 0488.

Различные методы надежной кластеризации имеют разные неявные способы классификации точек как «шум» (шумовая составляющая, обрезка, идентификация выбросов в t- распределениях). Делать их сопоставимы, мы используем (4.1) для унификации начисления баллов методы расчета $\mathrm{AGR}_\alpha(\cdot)$ на основе параметров при фиксированном квантиле $1-\alpha$ χ^2_2 (m_j, S_j) определяет эллипсоид в R^p, который размещает j- й кластер. Размер этого эллипсоида определяется в терминах

оценивается методами, из которых мы предполагаем, что оценка-торс троек (π_j, m_j, S_j) (пропорция кластера, центр и

Стр.7

ковариационная матрица) может быть вычислена (см. раздел 5.2). Пусть расчетные параметры кластера быть $\theta_{C,n}$. Коэффициент ошибочной классификации может затем вычисляется путем применения оптимальной перестановки $\sigma\{\cdot\}$ метки кластера:

$$\mathrm{MCR}(\theta_C, \theta_{C,N}) = \operatorname{Arg\,мин}_{\sigma} \frac{1}{n} \sum_{я=1}^{n} \mathbf{1}\{\mathrm{AGR}_{\alpha}(x_i; \theta_C) = \Sigma\{\mathrm{CMA}_{\alpha}(x_я; \theta_{C,N})\}\}. \qquad (4,2)$$

См. Онлайн-приложение для вычисления функции MCD. для ненормальных распределений.

5. Сравнительное моделирование.

Здесь мы представляем комплексное исследование с помощью моделирования OTRIMLE методами, представленными в разделе 2.

5.1. Процессы генерации данных

В общей сложности методы сравниваются на 24 DGP с 1000 копий Монте-Карло каждая. Половина DGP производит двумерные наборы данных. Остальные 12 DGP - 20-размерные версии, которые построены с добавлением независимых 18-мерная некоррелированная единичная дисперсия с нулевым средним Gaussian и / или студентов - m маргинальные. Поэтому кластеры всегда определяется только на первых двух маргиналах. Обратите внимание, что цель имитационное исследование - это не выбор переменных; мы разработали DGP, чтобы информация о кластеризации была только в первых двух измерения, чтобы иметь возможность визуализировать и контролировать схему кластеризации terns, но мы сравниваем методы кластеризации, использующие все переменные (использование методов выбора переменных выходит за рамки данного статья). Мы не думаем, что выбор или измерение переменных сокращение обязательно при кластеризации, потому что значение кластеры определяются задействованными переменными, см. ответ на обсуждение в Hennig and Liao (2013). Мы выбрали $n = 1000$ для двумерных конструкций и $n = 2000$ для 20-мерные версии. DGP были разработаны для тестирования разнообразие «шумовых паттернов», количества кластеров G и паттернов разделения / перекрытия между различными кластерами.

Мы рассматриваем два основных класса DGP, а именно DGP с равномерная составляющая шума на первых двух маргиналах и DGP которые не имеют шумовой составляющей. В первую группу входят следующие установки, все из которых имеют кластеры, сгенерированные из Gaussian распределений, а для $p = 20$ 18 неинформативных переменных являются гауссовыми: (i) для DGP с широким шумом равномерный шум составляет компонент дает точки, которые широко распространены, но перекрываются кластерные регионы целиком; (ii) для DGP «SideNoise» униформа шумовая составляющая распространяет точки на обширную область, выходящую за пределы слегка нахлестывается некоторыми гроздьями; (iii) в DGP «SunSpot» есть однородный компонент, который производит очень мало лежачие точки. С другой стороны, мы рассматриваем DGP, которые не включать шумовую составляющую (т. е. $\pi_0 = 0$). Эта вторая группа можно разделить на еще три подгруппы: (i) / (ii) в «GaussT» и «TGauss» DGPS, многомерный Студент - m распределения с используются три степени свободы. В «GaussT» они используются как неинформативные распределения для $p = 20$, тогда как первые два сгруппированных измерения используют гауссианы; в «TGauss» кластеры-тера порождаются нецентральными многомерными t_3 -распределениями и для $p = 20$ 18 неинформативных переменных являются гауссовскими; (iii) в «Бесшумных» DGP все точки взяты из гауссовского раздачи.

Для каждой из шести настроек есть варианты с нижним и нижним большее количество кластеров G , $p = 2$ (обозначается «l») и $p = 20$ (обозначается «h»), добавляя до 24 DGP. Используемая номенклатура в дальнейшем помещает их в конец имени установки, то есть «TGauss.5h» относится к настройке «TGauss» с более высоким $G = 5$ и более высокий $P = 20$. Для «WideNoise», «SideNoise» и «GaussT» нижняя G составляла 2, а большая G была 3. Для «SunSpot» «TGauss» и «Бесшумный»: нижняя G равнялась 3, а верхняя G - 5. перекрытие между кластерами, а также комбинации кластеров формы варьировались между DGP. Полная информация об определении DGP приведены в онлайн-приложении вместе с примерами диаграммы рассеяния первых двух измерений набора данных из каждую настройку.

5.2. Реализация методов

В таблице 1 приведены настройки сравниваемых методов. TCLUST и RIMLE / OTRIMLE основаны на собственном соотношении напряжения, но это не относится к программному обеспечению MCLUST (Fraley и другие. 2012 г.) и доступная реализация смесей t - распределения (McLachlan and Peel 2000). Чтобы иметь полную сравнительную В качестве решения можно использовать ограничения на собственные отношения (см. раздел 3.3) реализованы нами для OTRIMLE / RIMLE, gmix, tmix и gmix.u; последний вычисляется с использованием того же процедура, которая используется для RIMLE / OTRIMLE. Для TCLUST ограничения находятся в исходном R-пакете.

Для всех методов ограничение на собственное отношение установлено равным до 20 для каждого из 24 DGP. Последний выбор мотивирован тем, что 20 больше, чем максимальное истинное собственное отношение по проектам, участвующим в сравнении, и в общем Выберите значение, которое часто обеспечивает довольно плавную оптимизацию (очевидно, на самом деле истинные отношения собственных значений неизвестны, но для переменных со сравнимыми шкалами и диапазонами значений, 20 дает ковариационные матрицы достаточно гибкие для большинства приложений). OTRIMLE протестирован со штрафным сроком и без него. $\beta = \frac{1}{3}$ в (3,12), обозначаемые ot.rimle ($\beta = 0$) и ot.rimle.p, соответственно. Другие значения β в диапазоне от 0,1 до 0,5 имеют были опробованы, и результаты не сильно изменились Трудность с TCLUST - это автоматический выбор триммера на основе данных. уровень мин в настоящее время недоступен. В tclust.f устанавливаем trim-уровень мин до 10%. Этот выбор мотивирован тем, что DGP произвела среднюю долю точек, принадлежащих NR $_\alpha$ задается в диапазоне [0%, 23%] (см. Таблицу 1 у Коретто и Хеннига в

Таблица □. Сводка основных настроек сравниваемых методов.

Метод	Настраивать
gmix.u	RIMLE с $\delta = 1 / V_n$, где V_n объем наименьшего гипер прямоугольник, содержащий данные.
от.римле	OTRIMLE без штрафа ($\beta = 0$).
от.римле.п	OTRIMLE со сроком штрафной $\beta = 1 / 3$.
tclust	TCLUST с фиксированным уровнем обрезки, установленным на □.□.
ot.tclust	TCLUST с уровнем обрезки, выбранным критерием OTRIMLE без штрафа ($\beta = 0$).
ot.tclust.p	TCLUST с уровнем обрезки, выбранным критерием OTRIMLE с пенальти термин $p = 1 / 3$.
tmix	ML для Студент - m модели смеси (□.□) с $\nu = 3$ для всех компоненты плюс ограничения на собственное отношение.
gmix	ML для модели гауссовой смеси плюс ограничения на собственное отношение.

Стр. 8

нажмите b). Кроме того, поскольку уровень обрезки играет аналогичную роль lar к δ в RIMLE / OTRIMLE есть две версии TCLUST для чего та же идея для автоматических решений обрезка уровень используется, как предлагается здесь для OTRIMLE, см. Раздел 3. В ot.tclust и ot.tclust.p, уровень обрезки выбран с использованием (3,12) с уровнем обрезки, играющим роль δ , и

стандартные ошибки приведены в таблице 2.. Каждый квадрат в сюжете представляет собой цветовое представление степени ошибочной классификации усредненные по 1000 повторений Монте-Карло для данного пара метод-DGP. Дополнительные сведения о средней ошибочной классификации-ставки катионов приведены в онлайн-приложении. Он также кон-содержит коробчатые диаграммы уровней ошибочной классификации для всех методов-DGP веса $\tau_j(\cdot)$ в $M_j(\cdot)$ заменены на 0–1 четкие веса TCLUST, снова используя $\beta = 0$ или $\beta = \frac{1}{3}$. Для t- смесей в оригинальное предложение Маклахлана и Пила (2000 г.), степени свобода считается равной по составу смеси. значений и оцениваются по данным и ковариационной матрице ограничения не рассматриваются. В tmix фиксируем степени свобода до 3, и мы включаем ограничения на собственное отношение, как для другие методы. Это мотивировано следующим образом: (i) для некоторых из ограничения DGP (особенно конструкции SunSpot) были необходимо избегать ложных решений; (ii) для некоторых выборок конструкции не основывается на студентах - m распределения, Оценка степени свободы t -распределения дали чрезвычайно большой разброс показателей ошибочной классификации; (iii) поскольку число- бер конструкций основаны на студентов - m с 3 - мя степенями свободы, решение установить этот параметр на 3 дает t- смеси небольшое преимущество, которое кажется справедливым, учитывая, что большинство настроек скорее, похоже, предпочитает шумовой компонент / обрезку.

Опыт подсказывает, что для некоторых DGP решение Настройки могут сильно зависеть от инициализации. Чтобы уменьшить смещения, вызванного различными инициализациями, все методы являются [illegible] инициализированы из того же раздела, см. раздел 3.3 (дополнительная набор функций R был предоставлен авторами TCLUST для учтите это).

5.3. Полученные результаты

Методы сравниваются с использованием коэффициентов ошибочной классификации. определено в (4.2). Это более актуально в задачах кластеризации чем оценки параметров. Результаты представлены графически. показано на рисунке 2, в то время как средние показатели ошибочной классификации с пары. На рисунке 2 показаны четкие доказательства того, что использование надежных методов это важно. Метод gmix хорошо работает только с шумом. less и некоторые DGP с гауссианами и t -распределениями, но большинство других методов работают хорошо (хотя временами немного хуже) для этих DGP тоже. tmix хорошо работает для большинства задействованных DGP. ing t -распределения, но для других DGP с шумом / выбросами, часто он серьезно хуже, чем gmix.u, OTRIMLE и TCLUST.

gmix.u работает относительно хорошо, хотя для некоторых среди DGP сильно страдает большой размерностью. Для WideNose в двух измерениях, gmix.u во многих случаях будет соответствовать правильный оценщик ML, поэтому метод должен быть выгодным здесь, и gmix.u действительно лучше всего подходит для этих DGP. Тем не мение, для 20-мерного WideNoise, принимая равномерное распределение построение по наименьшему гипер прямоугольнику, содержащему данные [illegible] не может быть оценкой ML для генерирующего шум компонент смеси, и его характеристики ухудшаются сильно.

Что касается методов TCLUST, хотя автозаполнение matic уровень обрезки ot.tclust и ot.tclust.p не всегда улучшить результаты, продемонстрировав разумную выбор уровня обрезки. Фактически, для всех ситуаций, когда истинная средняя доля шума примерно равна обрезке уровень tclust, производительность tclust, ot.tclust и ot.tclust.p очень похожи, что означает, что критерий OTRIMLE является хорошая отправная точка для определения скорости обрезки. По сравнению с методы типа RIMLE, производительность TCLUST страдает в ситуациях, когда есть значительная степень перекрытия между кластерами. Для DGP с перекрытием (например, WideNoise.2, SunSpot.5, GaussT.2 и Noiseless.5), ошибочная классификация

Рисунок □. График уровня, представляющий среднее значение выборки распределения Монте-Карло вероятностей ошибочной классификации (процентная шкала) для каждой пары DGP-метод. Каждый квадрат график представляет собой среднее значение ошибочной классификации в соответствии с нижней шкалой серого цвета.

Таблица □. Средние показатели ошибочной классификации (%) Монте-Карло со стандартными ошибками в скобках Коэффициенты ошибочной классификации рассчитываются как в (□.□). Обратите внимание, что оба средних (и стандартные ошибки) представлены в процентной шкале.

	Метод							
DGP	gmix.u	от.римле	от.римле.п	tclust	ot.tclust	ot.tclust.p	tmix	gmix
WideNoise.□l	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)
WideNoise.□h	□□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)
WideNoise.□l	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)
WideNoise.□h	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)
SideNoise.□l	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)
SideNoise.□h	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)
SideNoise.□l	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)
SideNoise.□h	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)
Sunspot.□l	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)
Солнечное пятно.□h	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)
Sunspot.□l	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□.□□ (□.□□)	□□.□□ (□.□□)	□.□□ (□.□□)
Солнечное пятно.□h	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)
TGauss.□l	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)
TGauss.□h	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)
TGauss.□l	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)
TGauss.□h	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)
GaussT.□l	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)
GaussT.□h	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)
GaussT.□l	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)
GaussT.□h	□□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□□.□□ (□.□□)	□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)
Бесшумный.□l	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)

Бесшумный. H	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)
Бесшумный.□l	□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□.□□ (□.□□)	□.□□ (□.□□)
Бесшумный. H	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)	□□.□□ (□.□□)

уровень TCLUST полностью определяется ошибочными классификациями между кластерами (чтобы увидеть это, рассмотрите Таблицы 2 и 4 в онлайн-добавка). Причина в том, что параметры TCLUST основанный на правдоподобии типа классификации, который полагается на отдельные между кластерами. TCLUST также, кажется, не терпит - съел большое количество студентов - *m* маргинальных GaussT. и GaussT.3h. TCLUST хорошо работает для ряда DGP и явно лучше всего в WideNoise.2h.

Методы OTRIMLE показывают очень хорошую общую производительность. форма. Они дают высокие показатели ошибочной классификации только для некоторые 20-мерные DGP, для которых все методы находятся в проблема (производительность GaussT.2h в целом плохая с даже tmix, лучший метод, производящий среднюю ошибку уровень классификации более 30%), и они лучше всего подходят для ряд DGP, в частности 20-мерный WideNoise и некоторые TGauss-DGP. Сравнение от.римле и ot.rimle.p смешивается (как между ot.tclust и ot.tclust.p), с $\beta = \frac{1}{3}$ улучшение ситуации явно для TGauss.2l и TGauss.3l (форма *t*- распределения способствует присвоить шуму слишком много точек; см. Таблицы 2 и 3 в онлайн-приложении, но значительно хуже для WideNoise. 3ч.

Сравнивая OTRIMLE с gmix.u, есть ряд DGP, для которых gmix.u имеет несколько меньшую ошибочную классификацию оцените один или оба от.римле и от.римле.п. Во всех этих DGP, все gmix.u, ot.rimle и ot.rimle.p в основном производят та же структура кластеризации, с разногласиями только по поводу классификация некоторых пограничных точек. Различия есть более существенный в настройках, в которых gmix.u хуже. Для WideNoise.2h, gmix.u в большинстве случаев не обнаруживает шума, так что один из кластеров состоит в основном из шума. Для широкоформатных Шум. 3ч, иногда весь или большой шум объединяется в кластер, с некоторым влиянием на структуру кластеризации. Для GaussT.3h, значительное количество выбросов интегрировано в кластеры.

5.4. Поведение D (δ)

Здесь мы исследуем поведение D *(δ)* как функции от *δ с* помощью Эксперименты Монте-Карло при различных DGP из [раздела 5](). Для каждого из этих DGP мы создали 100 независимых выборок. Мы вычислили D *(δ)* для сетки значений *δ,* взятых из интервала $[2.22 \times 10^{-308}, 1]$, добавив *δ* = 0. На [рисунке 3](), мы сообщаем Монте Карло усредняет ± стандартные ошибки для D *(δ)* для двух выбранных DGP: WideNoise.3h и GaussT.3h. Они определены в [разделе 5.1.](), оба с *p* = 20. Основное отличие состоит в том, что создается шум двумерным равномерным распределением в WideNoise.3h, тогда как в GaussT.3h измерения 3–20 от центра t_3 распределения, генерируя некоторые выбросы. [Рисунок 3]() показывает поведение D *(δ)* при log *(δ)*> −200. Для меньших значений *δ* (включая *δ* = 0) поведение кривых в основном соответствовало стат. На обоих графиках есть явный минимум, хотя для GaussT.3h этот минимум лежит на границе. Для WideNoise.3h критерий OTRIMLE имеет приятное выпуклое поведение вокруг своего минимума. В GaussT.3h в размерностях 3–20 распределение форма кластеров отклонилась от гауссовости за счет более тяжелых хвостов, хотя ядро распределения похоже на Гаус- сиан. D *(·)* затем применяет гауссову форму, назначая много указывает на шумовую составляющую. В результате $\pi_{0,n}$ становится большой, а оптимальное *δ* достигается в точке, где большее значение Значения *δ* не дают оценок параметров в пределах кон- напряженный набор больше (если ограничение не принудительно). В последнее является основанием для применения срока штрафа в ([3,12]()) (видеть также [Раздел 5.5]()). Для остальных 22 DGP [раздела 5.1](), мы нашел похожие шаблоны (в основном похожие на Widenoise.3h здесь). Еще одно наблюдение на [рисунке 3]() заключается в том, что около минимального

Стр.10

Рисунок □. Среднее значение Монте-Карло для критерия OTRIMLE *D (δ)* (синяя сплошная линия) ± стандартные ошибки (пунктирные линии), вычисленное по сетке значений для *δ* ∈ {0} ∪ [2 . 22 × 10^−308, 1]. Эти два графика относятся к DGP «WideNoise.□h» и «GaussT.□h» в [Разделе □ .]()

D (δ) кажется достаточно стабильным для разных наборов данных из одного и того же дизайн.

5.5. Эффект β

Для всех 24 DGP, рассмотренных в [разделе 5.1](), мы также исследовали поведение доли шума $\pi_{0,n}$ в зависимости от *β* , видеть ([3,12]()). Для каждой конструкции было изготовлено 100 независимых образцов. элементов, и для каждого из них мы вычислили решение OTRIMLE. ции $\theta_n(\delta_n)$ для различных значений *p* ∈ [0 , 1]. [Рисунок 4]() отчеты среднее Монте-Карло оценочной доли шума $\pi_{0,n}$ ± стандартная ошибка. Как для WideNoise.3h, так и для GaussT.3h, увеличение *β* плавно снижает расчетную пропорцию шума. тион. Однако есть некоторая разница в масштабах. Влияние из *p* гораздо сильнее GaussT.3h, и то же самое происходит для всех тех планов выборки, в которых распределение внутри кластера ции отклоняются от гауссовости. Результаты для других DGP вполне удовлетворительны. похожий.

потому что при *β* = 0 только довольно маленькие центральные ядра (60–70%) точки, порожденные *t* -распределениями, были отнесены к кластерам ters, тогда как для больших *β* только точки были классифицированы как шум, который действительно были весьма «отдаленными». С другой стороны, при большем *β* в данные из DGP с гауссовыми кластерами и некоторым шумом, который был не четко отделены от кластеров, больше «истинных» шумовых точек были отнесены к кластерам. Нет объективного правила для чего следует учитывать процент точек от *t*- распределения. как «действительно» удаленный, и поэтому это должно быть решено пользователь, насколько «терпимым» OTRIMLE должен быть по отношению к более тяжелому распределительные хвосты, чем гауссовы.

[На рис.5]() показана ситуация, когда выбор *β* влияет на кластеризация много. Набор данных состоит из 100 наблюдений, каждое из которых *N (0 , 1)* и *N (3 , 1)* по оси *x* и 12 наблюдений из *N (12 , 25)* . OTRIMLE был оснащен *G* = 2. Первые два смешанных Компоненты не очень хорошо разделены. *β* = 0 не штрафовать за шум, и поэтому наблюдения из третьего компонент объявлен «шумовым», а первые два компонента

Открытие общей структуры кластеризации никогда не было зависит от изменения β от 0 до 0,5 для DGP с Раздел 5. В большинстве случаев изменений не было. кластеризация вообще. Единственное отличие заключалось в том , что больше β сома раз дали более низкий процент данных, классифицируемых как шум. В случае использования t_3 -распределений это было хорошо,

разделены. Однако $\beta = \frac{1}{3}$ объединяет первые два компонента а третий объявляет вторым кластером. «Переключение точка »между этими двумя способами« интерпретации »кластеризации структура находится примерно при $\beta = 0.3$; большие значения β не изменяют кластеризации больше. Несмотря на то, что «истинная» G здесь равна 3, с учетом толкование, в зависимости от приложения это вполне может имеет смысл трактовать наименьший компонент смеси как

Рисунок . Среднее Монте-Карло для оцененной доли шума $\pi_{0,n}$ (синяя сплошная линия) ± стандартные ошибки (пунктирные линии), вычисленных по сетке значений для $\beta \in [0, 1]$. В два графика относятся к DGP, названным «WideNoise.h» и «GaussT.h» в Разделе .

Стр. 11

Рисунок . Кластеры с $\beta = 0$ и $\beta = \frac{1}{3}$ на примере искусственных данных.

шум / выбросы, или объединить первые два компонента в один кластер. β настраивает метод на создание шума или, скорее, на допускать не очень гауссовские компоненты в подобных случаях.

6. Приложения

В этом разделе мы применяем OTRIMLE и альтернативные методы. упомянутые ранее к двум реальным наборам данных. В первом примере не содержит никакой «основной истины», тогда как второй «Настоящие» классы.

6.1. Дортмунд Данные

Здесь мы анализируем набор данных, дающий информацию о 170 дисквартала немецкого города Дортмунд, о котором рассказывается в Sommerer и Weihs (2005 г.). Мы использовали пять социологических ключевые переменные и преобразовали их таким образом, чтобы Имеет смысл рассматривать гауссовские распределения внутри кластеров. В результирующие переменные - это логарифм уровня безработицы («Безработица»), соотношение рождений / смертей, деленное на число жителей («рождение. смерть»), сальдо миграции, деленное на количество жителей («move.in.out»), логарифм ставка сотрудников, выплачивающих социальное страхование («soc.ins.emp»), и корень квадратный из числа жителей («жителей»).

Все переменные были центрированы и стандартизированы среднее абсолютное отклонение. На рисунке 6 показана диаграмма рассеяния рождение. смерть и ходы. вов. Чтобы справиться с заговором-тин, существующий крайний выброс со значениями ≈ $(-200, 50)$ является не показано. На рисунке 7 представлена диаграмма рассеяния безработицы. и soc.ins.emp. На рисунке 7 показаны более умеренные выбросы. Левая часть рисунка 6 показывает кластеризацию из подгонки простая гауссовская смесь с $G = 4$ для всех пяти переменных по Пакет MCLUST от R. Кластер 4 - это одноточечный кластер, состоящий из крайний выброс. Кластеры 1 и 3 в основном подходят для двух разных разновидностей умеренных выбросов, тогда как всего более 120 районы, которые не являются крайними по этим двум переменным объединены в единый кластер. Понятно, было бы больше желательно иметь кластеризацию, в которой не так сильно доминирует несколько странных районов, учитывая, что есть какая-то значимая структура среди других районов. Такая кластеризация производится Метод OTRIMLE, показанный на правой стороне рисунка 6 и на левая часть рисунка 7. Кластеризация хорошо интерпретируется с кластером нет. 3 сбор группы районов с высшим сальдо миграции и очень разбросанный коэффициент рождаемости / смертности, кластерный нет. 1 представляет собой кластер с высокой вариабельностью, характеризующийся высокой занятость или большое количество сотрудников, выплачивающих социальное страхование, кластер нет. 2 - однородная группа со средним числом сотрудников, выплачивающих социальное страхование, и довольно высокие, но не очень высокий уровень безработицы, а кластера нет. 4 собирает больше всего

Рисунок □. Диаграмма рассеяния рождения.смерти и перемещений.выход из набора данных Дортмунда с кластеризацией MCLUST (слева) и кластеризацией OTRIMLE (справа) с $G = 4$.

Стр. 12

Рисунок □. Диаграмма рассеяния soc.ins.emp и безработицы из набора данных Дортмунда с кластеризацией OTRIMLE (слева) и кластеризацией TCLUST со скоростью усечения □% (справа) с $G = 4$.

районы с низкими значениями обеих этих переменных. За это набор данных, значения от $\beta = 0$ до $\beta = 0.5$ дают то же самое кластеризация (с ограничением на собственное отношение $\gamma = 20$), несмотря на то, что $\beta = 0$ приводит к $\log(\delta) = -11.5$ и $\pi_{0,n} = 0.055$, тогда как при $\beta = 0.5$, (3,12) дает $\log(\delta) = -11.9$ и $\pi_{0,n} = 0.054$.

Рассматривая методы, представленные в разделе 2, оказывается что существенные разногласия могут существовать между разными надежные методы кластеризации. Применение реализованных методов в tclust и обсуждается в García-Escudero et al. (2011 г.), $\alpha = 0.07$ оказался хорошей скоростью обрезки для TCLUST с $G = 4$ здесь. Полученная кластеризация сравнивается с OTRIMLE на рисунке 7. Хотя то, что было обрезано, почти идентичен «шуму» OTRIMLE, кластеризация несколько отличается ferent, с кластерами TCLUST нет. 1 и 4 в диапазоне область «высокого разброса, характеризующаяся высоким уровнем безработицы и большое количество сотрудников, выплачивающих социальное страхование », представлен кластером нет. 3 TCLUST и кластер нет. 1 из ОТРИМЛЕ. Трудно интерпретировать номер кластера OTRIMLE. 4 с использованием любая пара переменных. Это можно увидеть в онлайн-приложении. (Коретто и Хенниг в прессе b), а также решения других методы.

6.2. Данные народных песен

Второй набор данных предоставил Даниэль Мюлленсифен. В наблюдений 776 народных песенных мелодий, из них 586 из Люксембурга, а остальные 190 из Вармии в Польша. Это «настоящие» классы. Мелодии оригинальны из базы данных мелодий ESAC (Schaffrath 1992). 18 февраля (см. онлайн-приложение (Коретто и Хенниг в прессе) б) для списка) рассчитывались программой «ФАНТАСТИЧЕСКИЙ» (Müllensiefen 2009 г.).

Визуальный осмотр обнаруживает много необычных мелодии, то есть выбросы в наборе данных. Основные объемы мелодии из Люксембурга и Вармии систематически отличаются друг от друга, хотя есть много перекрытий и нет сильных разделение. Для измерения того, в какой степени вычисляются кластеризации с $G = 2$ совпадали с двумя областями, использовались скорректированные Индекс Рэнда (ARI; Hubert, Arabie 1985) с ожидаемым значение 0 для двух случайных кластеров и максимум 1 для идеальное согласие. OTRIMLE с настройками, как указано выше ($\beta = 0$, $\gamma = 20$) классифицировал 36,9% наблюдений как «шум». В ARI между решением OTRIMLE и исходными регионами составляет 0,155. Для этого (как и для других методов кластеризации) Решение OTRIMLE интерпретировалось как трехкластерное решение. с «шумом» в качестве третьего кластера. MCLUST по умолчанию дает ARI = -0 . 045, MCLUST с шумом дает ARI = −0 . 017, от.ткласт дает ARI = 0 . 016 (оригинальная функция TCLUST с обрезкой-коэффициент ming 0,369, как предложено OTRIMLE выше, достигает ARI = 0 . 089), а tmix дает ARI = 0 . 083 . ARI-значение OTRIMLE, хотя явно лучше, чем у других методов, не является пар-особенно высокий, но вычисление ARI только по наблюдениям которые не были классифицированы как шум, достигает ARI = 0 . 392 (решение ция с $\beta = \frac{1}{3}$ здесь немного хуже, но чуть лучше выше относительно ARI, включая точки шума), что предполагает что существует четкое соответствие между кластерами OTRIMLE - регионов теринг и мелодии.

7. Заключительные замечания

Несмотря на наши усилия сделать исследование моделирования справедливым, в конечном итоге было бы хорошо иметь сравнения методов, выполняемых исследователи, не принимавшие участия в разработке каких-либо методов. Каждый метод лучше всего подходил для определенных DGP в исследование моделирования, и исследования моделирования могут быть разработаны которые делают любой метод «выигрышным». Читатели должны составить свой собственное мнение о том, в какой степени наше исследование охватило ситуации что для них важно. Одна из наших основных целей заключалась в том, чтобы перед всеми методами DGP, которые не совсем соответствуют их предположения модели, но для которых методы, тем не менее могут быть использованы на законных основаниях. Фактически, мы воплотили важные идеи как из MCLUST (инициализация), так и из TCLUST (собственное значение соотношения ограничений), и комбинация этих идей, используемых здесь действительно может быть полезным для всех методов. Методы предварительного Присланные в этой статье скоро будут доступны в новом R-пакете ОТРИМЛЕ.

Задача определения количества кластеров G такова: очень важно на практике. Вот два возможных подхода для RIMLE. Во-первых, самый популярный подход к подгонке простой смеси - модели, а именно байесовский информационный критерий, могут использоваться (обработка RIMLE / OTRIMLE неправильной постоянной плотности как правильный), Фрейли и Рафтери (1998 г.). Во-вторых, G мог

Стр. 13

решаться на исследовательской основе, отслеживая изменения псевдо-правдоподобие по разным значениям G аналогичным образом тому, что сделано для TCLUST в Гарсиа-Эскудеро и др. (2011 г.). Глубокое исследование этих подходов выходит за рамки эта статья.

Дополнительные материалы

Дополнительные материалы содержат: (i) дополнительную информацию о методах и алгоритмы; (ii) подробные определения планов выборки для моделирования исследование вместе с коробчатыми диаграммами и сводными таблицами для неправильной классификации ставки (iii) дополнительные графики для приложений с реальными данными; (iv) Программное обеспечение R с инструкции по воспроизведению сравнительного моделирования и применения к реальным данным.

Финансирование

Авторы выражают благодарность за финансовую поддержку исследованиям МИУР. грант PRIN 2010J3LZEN, а также использование высокопроизводительной ком- строительство инфраструктуры финансируется Университетом Салерно (исследовательская программа ASSA098434). Авторы выражают благодарность за поддержку Грант EPSRC EP / K033972 / 1.

ORCID

Пьетро Коретто http://orcid.org/0000-0002-6972-9671
Кристиан Хенниг http://orcid.org/0000-0003-1550-5637

Рекомендации

Банфилд, Дж. Д., и Рэфтери, А. Э. (1993), «Гауссовские и нестандартные модели. Гауссовская кластеризация, *Биометрия* , 49, 803–821. [1648 , 1649, 1651 г.]
Беккер, К., и Гатер, У. (1999), «Точка разрушения маскировки в мульти- различные правила идентификации выбросов », *Journal of the American Statistical Ассоциация* , 94, 947–955. [1652]
Байерс, С., и Рэфтери, А.Е. (1998), «Удаление помех от ближайшего соседа для Оценка характеристик в процессах пространственных точек », *Американский журнал. Статистическая ассоциация* , 93, 577–584. [1651]
Катор, ЕА, и Lopuhaä, HP (2012), "Центральная предельная теорема и влияние Функция оценки MCD на общем многомерном распределении бюллетеней », *Бернулли* , 18, 520–551. [1652]
Коретто, П., и Хенниг, К. (2010), «Исследование моделирования для сравнения надежных Методы кластеризации на основе смесей »,« *Достижения в области анализа данных и Классификация* , 4, 111–135. [1648]
——— (в печати), «Последовательная и устойчивая к отказу модель- метод кластеризации на основе », доступный по адресу *http://arxiv.org/pdf/1309.6895* . [1649, 1650 , 1651]
——— (в печати b), Дополнение к: «Устойчивый несоответствующий максимум. Вероятность: настройка, вычисление и сравнение с Другие методы для робастной гауссовской кластеризации », доступный по адресу *http://arxiv.org/src/1406.0808/anc/supplement.pdf* . [1649, 1654 , 1658]
Куэста-Альбертос, JA, Gordaliza, A., и Matrán, C. (1997), «Trimmed k- означает: попытка робастизации квантователей », *Annals of Statistics* , 25, 553–576. [1649]
Croux, C., и Haesbroeck, G. (1999), «Функция влияния и эффективность. оценки матрицы рассеяния, определяющей минимальную ковариацию », *Журнал многомерного анализа* , 71, 161–190. [1652]
Дэвис, Л., и Гэзер, У. (1993), «Идентификация множественных выбросов», *Журнал Американской статистической ассоциации* , 88, 782–792. [1652]
Дэй, штат Нью-Йорк (1969), «Оценка компонентов смеси нормальных Распределения, *Биометрика* , 56, 463–474. [1650]
Демпстер, А.П., Лэрд, Н.М., и Рубин, Д.Б. (1977), «Максимально подобное - достоверность неполных данных с помощью алгоритма ЭМ », *Журнал Королевское статистическое общество* , Series B, 39, 1–47. [1649]
Деннис, JEJ (редактор) (1981), *Алгоритмы для нелинейной подгонки, (НАТО Симпозиум перспективных исследований)* , Кембридж, Великобритания: Cambridge University Press. [1650]
Фрейли, С., и Рэфтери, А.Е. (1998), «Сколько кластеров? Какой кластер- метод теринга? Ответы через модельно-ориентированный кластерный анализ », *The Computer Journal* , 41, 578–588. [1658]
Фрейли, К., Рафтери, А.Е., Мерфи, ТБ, и Скракка, Л. (2012), *mclust Версия 4 для R: Моделирование нормальной смеси для кластеров на основе моделей. Теринг, классификация и оценка плотности* , Технический отчет No. 597, Статистический факультет Вашингтонского университета, Сиэтл, Вашингтон. [1648 , 1649, 1653 г.]
Фриц, Х., Гарсия-Эскудеро, Лос-Анджелес, и Майо-Искар, А. (2012), «tclust: An R Пакет для триммингового подхода к кластерному анализу », *Journal of Sta- статистическое программное обеспечение* , 47, 1–26. [1649]
Гальегос, М. Т. (2002), «Кластеризация с максимальным правдоподобием с выбросами», в *Классификации, кластеризации и анализе данных* , ред. К. Яджуга, А. Соколовский, Х.-Х. Бок, Берлин: Springer, стр. 247–255. [1649]
Гальегос, М. Т., и Риттер, Г. (2005), «Надежный метод кластерного анализа. ysis, *Annals of Statistics* , 33, 347–380. [1649]
——— (2009 г.), «Оценка ограниченного ML загрязненных смесей», *Санкхья (Серия А)* , 71, 164–220. [1650]
Гарсия-Эскудеро, Л.А., Гордалиса, А., Матран, К., и Майо-Искар, А. (2008), «Общий подход обрезки к надежному кластерному анализу», *Анналы статистики* , 38, 1324–1345. [1649, 1650 г.]
——— (2010), «Обзор надежных методов кластеризации», « *Достижения в области данных» Анализ и классификация* , 4, 89–109. [1649]
——— (2011), «Изучение числа групп в надежных моделях. Кластеризация, *Статистика и вычисления* , 21, 585–599. [1658]
Hathaway, RJ (1985), «Ограниченная формулировка максимума - Оценка правдоподобия для нормальных распределений смеси », *Анналы. статистики* , 13, 795–800. [1650]
Хенниг, К. (2004), «Точки разрыва для оценок максимального правдоподобия. смесей местоположения и масштаба »,« *Анналы статистики »* , 32, 1313–1340. [1648 , 1649, 1650 г.]
——— (2010), «Методы слияния компонентов гауссовой смеси», *Достижения в области анализа и классификации данных* , 4, 3–34. [1651]
Хенниг, К., Ляо, Т.Ф. (2013), «Как найти подходящую кластеризацию. для переменных смешанного типа применительно к социально-экономическому расслоению. катион »(с обсуждением), *Journal of the Royal Statistical Science* , Series C, 62, 309–369. [1648, 1653 г.]
Хьюберт, Л., и Араби, П. (1985), «Сравнение разделов», *Journal of Classi- фикция* , 2, 193–218. [1658]
Ingrassia, S. (2004), «Ограниченный алгоритм, основанный на правдоподобии для нескольких трехмерные модели нормальной смеси », *Статистические методы и приложения* , 13, 151–166. [1650]
Инграссия, С., Роччи, Р. (2007), «Монотонные электромагнитные алгоритмы с ограничениями. для конечной смеси многомерных гауссианов, *Вычислительная статистика. и анализ данных* , 51, 5339–5351. [1650]
——— (2011), «Вырожденность алгоритма ЭМ для MLE многопрофильного варьировать гауссовские смеси и динамические ограничения », *Вычислительные Статистика и анализ данных* , 55, 1715–1725. [1650]
Кифер, Дж. (1953), «Последовательный поиск минимакса максимума», *Труды Американского математического общества* , 4, 502–506. [1651]
Маклахлан, Дж. Дж., И Пил, Д. (2000), «Робастное моделирование смеси с использованием t – распределение, *Статистика и вычисления* , 10, 339–348. [1649, 1653 г.]
Мюлленсифен, Д. (2009), *Фантастика: доступ к технологии анализа функций Статистика (в корпусе): технический отчет v1.5* , Лондон: Goldsmiths Лондонский университет. [1658]
Нейков, Н., Фильцмозер, П., Димова, Р., Нейтчев, П. (2007), «Robust Fit- тинг смесей с использованием усеченной оценки правдоподобия ", *Computa- Национальная статистика и анализ данных* , 17, 299–308. [1649]
Писон, Г., Элст, С.В., и Виллемс, Г. (2002), «Поправки для малых выборок. для LTS и MCD, *Метрика* , 55, с. 111–123. [1652]
Qin, Y., and Priebe, CE (2013), «Оценка максимального lq-правдоподобия. с помощью алгоритма максимизации ожидания: робастная оценка Модели смеси », *Журнал Американской статистической ассоциации* , 108, 914–928. [1649]
Руссеу, П.Дж. (1985), «Многомерное оценивание с высокой степенью разбивки. Point, *Математическая статистика и приложения* , 8, 283–297. [1652]
Шаффрат, Х. (1992), «Базы данных esac и программное обеспечение *Mappet* », *Comput- по музыковедению* , 8, 66. [1658]
Соммерер, Э.-О., и Вейз, К. (2005), «Введение в конкурс» Социальные Milieus in Dortmund »в книге« *Классификация - повсеместная проблема »* , Берлин: Springer, стр. 667–673. [1657]