Численный анализ неопределенности алгоритмов идентификации графовых моделей*

И.Д. Костылев, В.А. Калягин

Лаборатория алгоритмов и технологий анализа сетевых структур,
Национальный исследовательский университет
Высшая школа экономики в Нижнем Новгороде

Рассматривается известная в интеллектуальном анализе данных задача идентификации графа концентраций по наблюдениям в общей постановке. Исследуется неопределенность алгоритмов идентификации графа концентраций. Основное внимание уделяется анализу влияния плотности графа концентраций и выбора меры зависимости на неопределенность различных алгоритмов идентификации. В качестве алгоритмов идентификации используются алгоритмы множественной проверки гипотез. Реализован генератор случайных графовых моделей. Проведены массивные вычислительные эксперименты по оценке неопределенности идентификации графа концентраций в различных сетях случайных величин в зависимости от плотности графа. Сделаны выводы.

Ключевые слова: графовые вероятностные модели, Марковские случайные поля, граф концентраций, плотность графа, алгоритмы идентификации графа концентраций, неопределенность алгоритмов, множественная проверка гипотез.

1. Введение

Графовые модели стали популярным инструментом интеллектуального анализа данных в последние десятилетия. В работе рассматривается один из аспектов анализа вероятностных графовых моделей (Probabilistic Graphical Models or Random Markov Fields): идентификация графа концентраций по наблюдениям. Задача идентификации графа концентраций (Graphical Model Selection Problem) хорошо известна и активно исследуется в литературе (см. обзоры [1, 2]). Вместе с тем, в литературе пока еще недостаточно изучено влияние параметров графовой модели на качество алгоритмов идентификации графа концентраций. Ввиду высокой сложности теоретического анализа такого влияния, наиболее перспективным представляется подход, основанный на специально построенных вычислительных экспериментах. Для проведения вычислительных экспериментов требуется генератор случайных неотрицательно определенных матриц с заданной плотностью ненулевых элементов. В данной работе мы используем общую постановку задачи идентификации графа концентраций в сети случайных величин и методологию анализа влияние параметров графовой модели на качество алгоритмов идентификации, предложенную в нашей работе [3]. Основное отличие данной работы в использовании другого генератора случайных графовых моделей. В работе [3] использовался генератор случайных графовых моделей с заданной плотностью графа концентраций, основанный на разложении Холецкого. Целью настоящей работы является численный анализ неопределенности идентификации графовой модели в зависимости от плотности графа. В настоящей работе мы представляем результаты массивных вычислительных экспериментов по оценке влияния плотности графа на качество (неопределенность) алгоритмов идентификации графа концентраций с использованием генератора случайных графовых моделей, построенного на основе принципа доминирующей диагонали. Результаты проведенных экспериментов показывают близкое поведение неопределенности идентификации графа концентраций в зависимости от плотности графа для различных генераторов случайных графовых моделей. Это подтверждает адекватность предложенной методологии анализа и открывает новые перспективные задачи для исследо-

* Работа подготовлена в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ).

вания. В частности, для обоих генераторов заметно резкое ухудшение качества алгоритмов идентификации с ростом плотности графа. Этот феномен требует дополнительного исследования и модификации алгоритмов для повышения их качества.

Работа построена следующим образом. В разделе 2, следуя [3], мы кратко напоминаем постановку задачи, методологию оценки неопределенности алгоритмов идентификации графа концентраций и описываем рассматриваемые алгоритмы идентификации. В разделе 3 приведено описание генератора случайных графовых моделей, построенного на основе принципа доминирующей диагонали. В разделе 4 дано краткое описание результатов экспериментов и приведена ссылка на ресурс с полным представлением результатов. В последнем разделе приведены выводы по представленным материалам и описаны дальнейшие направления исследований.

2. Задача идентификации графа концентраций

Сетью случайных величин мы называем пару (X, γ), где $X = (X_1, X_2, \dots, X_N)$ случайный вектор размерности N, γ – мера зависимости между парой случайных величин. Предполагается, что значение меры зависимости $\gamma(X_i, X_j) = 0$ является признаком независимости случайных величин X_i, X_j. Сеть случайных величин порождает полный взвешенный граф Γ с N вершинами, в котором вес ребра (i, j) определяется величиной $\gamma_{i,j} = \gamma(X_i, X_j)$. Графом концентраций $Conc(\Gamma)$ графа Γ назовем граф с N вершинами, в котором ребро $(i, j) \in Conc(\Gamma)$ тогда и только тогда, когда $\gamma_{i,j} \neq 0$. Граф концентраций отражает графовую модель зависимостей компонент случайного вектора X, связанную с мерой зависимости γ. Классическая и наиболее популярная графовая модель связана с условными зависимостями компонент случайного вектора, что соответствует выбору меры частных корреляций, как меры зависимости. В настоящей работе рассматриваем граф концентраций в Гауссовских сетях случайных величин (вектор X имеет нормальное распределение) для корреляционных мер зависимости: классическая для этой задачи сеть частных корреляций (порождает Гауссовскую графическую модель, Gaussian Graphical Models), сеть корреляций Пирсона, сеть корреляций Фехнера и сеть корреляций Кендалла. Важно отметить, что для заданного случайного вектора X с многомерным нормальным распределением графы концентраций в сетях корреляций Пирсона, Фехнера и Кендалла совпадают. Этот верно и для более широкого класса эллиптических распределений [4].

2.1 Общая постановка задачи идентификации

Для заданной сети случайных величин (X, γ) и графа Γ Задача идентификации графа концентраций состоит в восстановлении графа $Conc(\Gamma)$ по наблюдениям. Под наблюдениями мы понимаем выборку размера n из распределения X. Моделью наблюдений является набор независимых одинаково распределенных случайных величин $X(1), X(2), \dots, X(n)$. Любой алгоритм идентификации графа концентраций можно рассматривать как отображение из выборочного пространства (sample space) в пространство неориентированных, невзвешенных простых графов (graph space). Неопределенность алгоритма идентификации связана с различием между графом $Conc(\Gamma)$ и графом, полученным в результате применения алгоритма идентификации. Качество алгоритма идентификации определяется численной характеристикой различия графов т.е. большим или малым значением неопределенности.

2.2 Методология оценки неопределенности

Мы рассматриваем задачу идентификации графа концентраций как задачу бинарной классификации ребер полного, неориентированного простого графа с N вершинами. Ребро (i, j) относим к классу "Yes", если оно присутствует в графе $Conc(\Gamma)$, и относим его к классу "No", если его нет в графе $Conc(\Gamma)$. Любой алгоритм идентификации графа концентраций в этом случае можно рассматривать как алгоритмом классификации. Для оценки качества алгоритмов бинарной классификации традиционно используется стандартная матрица ошибок со следующими элементами: TP (True Positive, число верно классифицированных ребер класса "Yes"), TN (True Negative, число верно классифицированных ребер класса "No"), FP (False Positive, число ребер

класса "No", ложно отнесенных к классу "Yes"), FN (False negative, число ребер класса "Yes", ложно отнесенных к классу "No"). Для оценки качества алгоритмов идентификации графа концентраций по наблюдениям (по выборке размера n из распределения случайного вектора X) мы используем математические ожидания (по распределению X) следующих характеристик качества: True Negative Rate TNR=TN/(TN+FP), True Positive Rate TPR=TP/(TP+FN), False Discovery Rate FDR=FP/(TP+FP), False Omission Rate FOR=FN/(TN+FN). Эти четыре характеристики отражают разные аспекты качества алгоритмов идентификации. Для оценки математических ожиданий характеристик качества мы используем среднее значение по заданному числу повторений экспериментов. В качестве обобщенных характеристик качества алгоритмов идентификации графа концентраций мы рассматриваем три характеристики: классические Balanced Accuracy BA, F1 score, и сравнительно недавно ставший популярным Matthew Correlation Coefficient MCC.

2.3 Алгоритмы идентификации

В работе исследуются алгоритмы идентификации графа концентраций, основанные на множественной проверке гипотез. Задача идентификации графа концентраций в постановке множественной проверки гипотез может быть выражена следующим образом:

$$\mathrm{H}_{ij}: \gamma_{ij} = 0 \text{ против } K_{ij}: \gamma_{ij} \neq 0, \qquad i, j = 1, \dots, N.$$

Статистические тесты используются для определения нулевых значений попарной меры связи между X_i и X_j на заданном уровнем значимости α. Если гипотеза была отвергнута, ребро (i, j) включается в множество ребер $Conc(\Gamma)$. В противном случае соответствующее ребро не включается в множество ребер. Для полной идентификации графа концентраций необходимо проверить C_N^2 гипотез, т.е. проверить каждую пару компонент случайного вектора X.

Для каждого конкретного вида сети случайных величин необходимы разные тестовые статистики. Детальное описание каждой из используемых тестовых статистик может быть найдено в [4].

Один тест контролирует вероятность ошибки первого рода на уровне значимости α. При этом при совместном применении всех индивидуальных тестов не контролирует FWER (вероятность допустить хотя бы одну ошибку первого рода на всем множестве из $M = \mathrm{C}_N^2$ тестов). Для решения данной проблемы используются различные поправки на множественность гипотез, контролирующие FWER или FDR на заданном уровне значимости. В работе рассматриваются следующие алгоритмы множественной проверки гипотез:

Simultaneous Inference (SI): Алгоритм множественной проверки гипотез без контроля FWER и FDR.

1. Вычислить тестовые статистики T_{ij} и вычислить соответствующие им p-значения p_{ij}.
2. Сравнить полученные p-значения с выбранным уровнем значимости. Если $p_{ij} < \alpha$, ребро $(i, j) \in E$, иначе ребро не включается в множество ребер.

Bonferroni correction (B): Алгоритм множественной проверки гипотез с контролем FWER.

1. Вычислить тестовые статистики T_{ij} и вычислить соответствующие им p-значения p_{ij}.
2. Вычислить количество тестов $M = C_N^2$.
3. Сравнить полученные p-значения с $\frac{\alpha}{M}$. Если $p_{ij} < \frac{\alpha}{M}$, ребро $(i, j) \in E$, иначе ребро не включается в множество ребер.

Holm step down procedure (H): Алгоритм множественной проверки гипотез с контролем FWER.

1. Вычислить тестовые статистики T_{ij} и вычислить соответствующие им p-значения p_{ij}.
2. Упорядочить полученные p-значения по возрастанию $p_1 \leq \cdots \leq p_M, M = C_N^2$.
3. Найти $R = \min\left\{k: p_k > \frac{\alpha}{M+1-k}\right\}$ и отклонить все гипотезы, соответствующие $k = 1, \dots, R-1$.
4. Для каждой отклоненной гипотезы добавить соответствующее ребро в множество ребер.

Benjamini-Hochberg procedure (BH): Алгоритм множественной проверки гипотез с контролем FDR для независимых тестовых статистик (не в случае задачи идентификации графа концентраций).

1. Вычислить тестовые статистики T_{ij} и вычислить соответствующие им p-значения p_{ij}.

2. Упорядочить полученные p-значения по возрастанию $p_1 \leq \cdots \leq p_M, M = C_N^2$.
3. Найти $R = \max\left\{k: p_k \leq \frac{\alpha k}{M}\right\}$ и отклонить все гипотезы, соответствующие $k = 1, \dots, R-1$.
4. Для каждой отклоненной гипотезы добавить соответствующее ребро в множество ребер.

Benjamini-Yekutieli procedure (BY): Алгоритм множественной проверки гипотез с контролем FDR в общем случае (даже если тестовые статистики не независимы)

1. Вычислить тестовые статистики T_{ij} и вычислить соответствующие им p-значения p_{ij}.
2. Упорядочить полученные p-значения по возрастанию $p_1 \leq \cdots \leq p_M, M = C_N^2$.
3. Найти $R = \max\left\{k: p_k \leq \frac{\alpha k}{M c_M}\right\}, c_M = \sum_{l=1}^{M} \frac{1}{l}$ и отклонить все гипотезы, соответствующие $k = 1, \dots, R-1$.
4. Для каждой отклоненной гипотезы добавить соответствующее ребро в множество ребер.

3. Генератор графовых моделей с заданной плотностью графа концентраций

Для оценки математического ожидания неопределенности алгоритмов идентификации графа концентраций необходим генератор случайных графовых моделей с заданной плотностью графа. Такой генератор должен создавать случайные положительно определенные матрицы с заданной плотностью ненулевых элементов.

В работе используется генератор матриц с доминирующей диагональю, предложенный в [5]. В начале работы генератор использует модель случайных графов Эрдеша-Реньи для создания случайной матрицы смежности с заданной плотностью. Затем ненулевым элементам матрицы присваиваются случайные веса из равномерного распределения на интервале $[-1, -0.5] \cup [0.5, 1]$, и каждый вес нормируется согласно правилу доминирующей диагонали. В конце происходит симметризация такой взвешенной матрицы смежности. Алгоритм создания случайной положительно определенной матрицы с заданной плотностью может быть описан следующим образом:

1. зафиксировать параметры N, p;
2. получить матрицу смежности A, соответствующую случайному графу Эрдеша—Реньи $G(N, p)$;
3. вычислить матрицу B, где $B_{ij} = A_{ij} * Uniform([-1, 0.5] \cup [0.5, 1])$;
4. преобразовать матрицу C: $C_{ij} = \frac{B_{ij}}{1{,}5 * \sum_k B_{ik}}$;
5. получить результирующую матрицу симметризацией матрицы C: $\frac{C + C^T}{2}$.

Представленный генератор был реализован на языке Python с использованием библиотек Numpy и NetworkX.

```
def generateDominantDiagonal(dim: int, density: float) -> tuple:
    graph = nx.gnp_random_graph(dim, density)
    adj = nx.adjacency_matrix(graph).toarray()

    A = np.random.uniform(0.5, 1, size=(dim, dim))
    B = np.random.choice([-1, 1], size=(dim, dim))
    prec = adj * A * B
    rowsums = np.sum(np.abs(prec), axis=1)
    rowsums[rowsums == 0] = 0.0001

    prec = prec / (1.5 * rowsums[:, None])
    prec = (prec + prec.T) / 2 + np.eye(dim)

    precision = prec
    covariance = np.linalg.inv(precision)

    pD = np.diag(1 / np.sqrt(np.diag(precision)))
    pcorr = -(pD @ precision @ pD)
    np.fill_diagonal(pcorr, 1)

    return covariance, precision, pcorr
```

4. Численный анализ неопределенности идентификации графовых моделей

4.1 Описание численных экспериментов

Численный анализ неопределенности алгоритмов идентификации графа концентраций требует проведения массивных экспериментов на семействе случайных графовых моделей, созданных при помощи генератора матриц с доминирующей диагональю. Параметры экспериментов:

- размерность случайного вектора N = 20;
- размер выборки n = 40, 100, 300;
- распределение данных: нормальное распределение с нулевым средним;
- количество случайных графовых моделей $S_{sg} = 500$;
- количество повторений для выбранной графовой модели $S_{obs} = 100$;
- значения плотности графа: $d \in [0.1, 0.9]$ (20 равноудаленных точек на интервале);
- уровень значимости индивидуальных тестов $\alpha = 0.05$.

Один эксперимент проводится по следующей схеме:

1. зафиксировать плотность d;
2. сгенерировать случайную графовую модель размерности N с заданной плотностью;
3. сгенерировать выборку размера n из распределения вектора X;
4. применить алгоритмы идентификации графа концентраций к выборке и вычислить значения метрик качества;
5. повторить шаги 3–4 S_{obs} раз и усреднить полученные значения метрик;
6. повторить шаги 2–5 S_{sg} раз и усреднить полученные значения метрик качества для оценки математического ожидания неопределенности алгоритмов на семействе случайных графовых моделей.

Выполнение всех шагов является достаточно трудоемкой вычислительной задачей. Для решения данной проблемы мы используем возможности суперкомпьютера для распараллеливания вычислений, выполняемых для фиксированных графовых моделей. Таким образом, каждая из S_{sg} графовых моделей обрабатывается независимо от остальных. Затем, после того как все графовые модели будут обработаны, итоговые метрики качества собираются и усредняются. Такой параллелизм позволяет значительно уменьшить время выполнения экспериментов.

4.2 Результаты численных экспериментов

Результаты численных экспериментов представлены в виде графиков зависимости от плотности характеристик неопределенности алгоритмов множественной проверки гипотез для идентификации графа концентраций в различных сетях случайных величин для нормального распределения случайного вектора *X*. В экспериментах использован генератор случайных графовых моделей, основанный на принципе доминирующей диагонали. Характерные результаты, отражающие результаты в целом, представлена на рис. 1–4. Полные результаты всех проведенных экспериментов можно найти по ссылке: https://github.com/cofofprom/gms_uncertainty/blob/main/Dominant%20Diagonal%20density%20experiments.pdf.

На рис. 1–4 выбраны классическая для этой задачи сеть частных корреляций и сеть корреляций Кендалла. На рисунках используются следующие сокращения для алгоритмов: SI (Simultaneous Inference), B (Bonferroni), H (Holm), BH (Benjamini-Hochberg), BY (Benjamini-Yekutieli). Как можно видеть из представленных графиков, все алгоритмы обеспечивают высокое качество характеристики TNR идентификации класса "No" (отсутствие ребра). Это соответствует известным теоретическим свойствам алгоритмов, связанных с гарантированным контролем ошибки FWER (вероятность хотя бы одного ложного ребра). Все алгоритмы, кроме SI успешно контролируют FDR, что тоже соответствует известным теоретическим результатам. Вместе с тем, характеристики качества идентификации класса "Yes" (присутствие ребра в графе концентраций) являются весьма далекими от приемлемых. Кроме того, явно виден рост ошибки

FOR (ложное отсутствие ребра) с ростом плотности графа концентраций. Обобщенные характеристики BA, F1, MCC качества алгоритмов множественной проверки гипотез для идентификации графа концентраций показывают низкое качество алгоритмов, особенно при росте плотности графа. Отмеченные особенности характерны для идентификации графа концентраций во всех рассматриваемых сетях случайных величин.

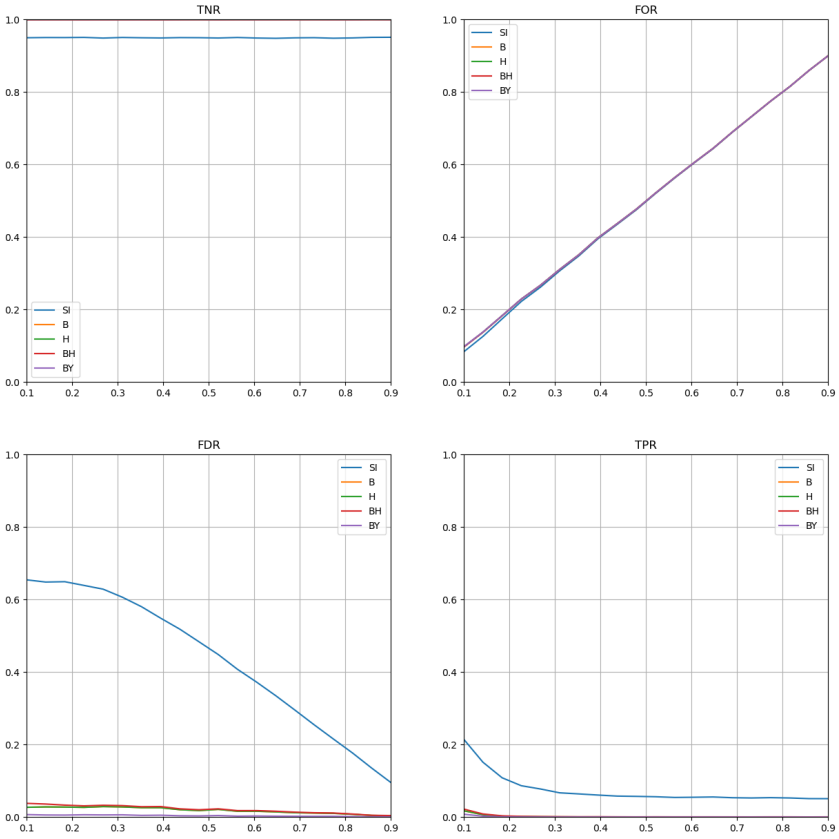

Рис. 1. Зависимость от плотности графа характеристик неопределенности TNR, FOR, FDR, TPR алгоритмов множественной проверки гипотез для идентификации графа концентраций в сети частных корреляций для нормального распределения вектора X и размера выборки $n=40$ (Gaussian Graphical Model Selection Problem)

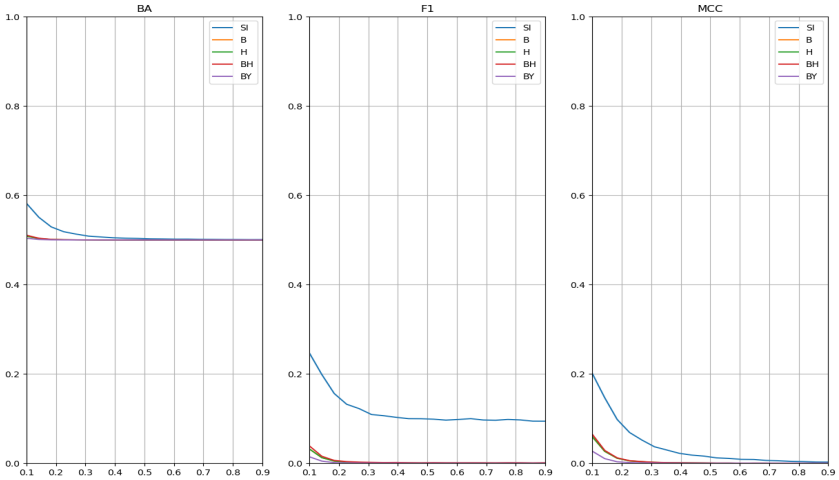

Рис. 2. Зависимость от плотности графа обобщенных характеристик неопределенности BA, F1, MCC алгоритмов множественной проверки гипотез для идентификации графа концентраций в сети частных корреляций для нормального распределения вектора X и размера выборки $n=40$ (Gaussian Graphical Model Selection Problem)

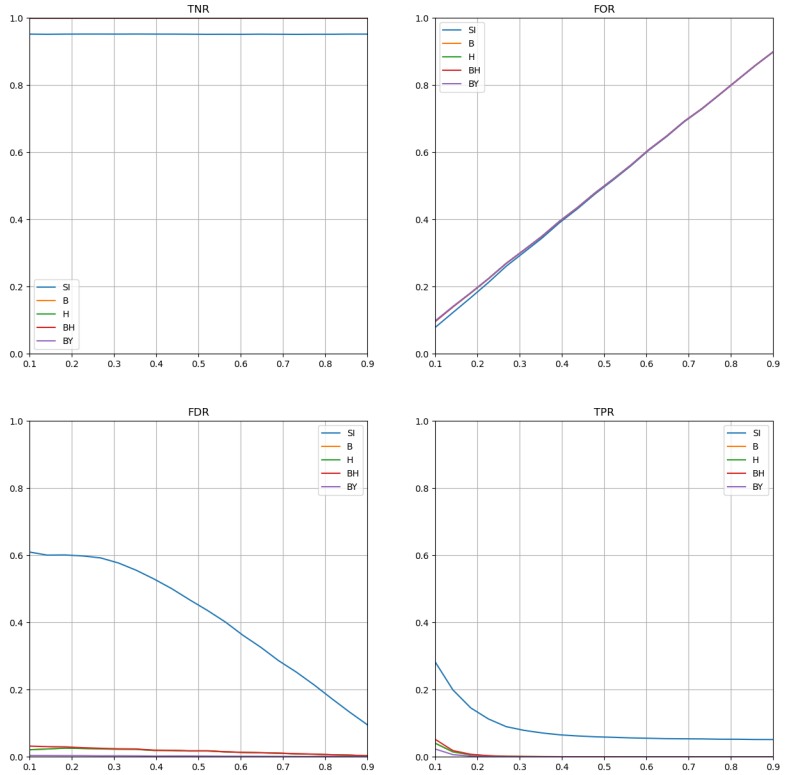

Рис. 3. Зависимость от плотности графа характеристик неопределенности TNR, FOR, FDR, TPR алгоритмов множественной проверки гипотез для идентификации графа концентраций в сети корреляций Кендалла для нормального распределения вектора X и размера выборки $n=40$ (Gaussian Graphical Model Selection Problem)

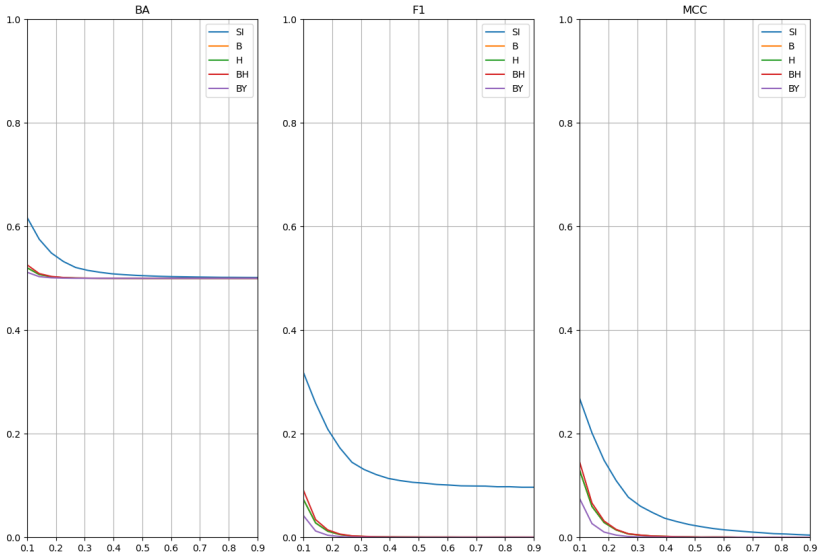

Рис. 4. Зависимость от плотности графа обобщенных характеристик неопределенности BA, F1, MCC алгоритмов множественной проверки гипотез для идентификации графа концентраций в сети корреляций Кендалла для нормального распределения вектора X и размера выборки $n=40$ (Gaussian Graphical Model Selection Problem)

5. Выводы

Настоящая работа является вкладом в исследование общей задачи анализа существующих и разработки новых алгоритмов идентификации графовых моделей в различных сетях случайных величин. В перспективе предполагается исследование устойчивости популярных алгоритмов и разработка новых устойчивых алгоритмов идентификации. Представленные вычислительные эксперименты направлены на изучение зависимости от плотности графа неопределенности алгоритмов множественной проверки гипотез для идентификации графа концентраций. Анализ результатов показывает, что независимо от выбранного генератора случайных графовых моделей общие тенденции зависимости от плотности качества алгоритмов множественной проверки гипотез совпадают и отражают специфику этих алгоритмов. В представленных результатах заметно резкое ухудшение качества алгоритмов идентификации с ростом плотности графа. Этот феномен требует дополнительного теоретического и экспериментального исследования для возможной адаптации алгоритмов множественной проверки гипотез при изменении плотности графа. Вычислительные эксперименты выполнены на суперкомпьютере cHARISMa НИУ ВШЭ в рамках проекта «графовые нейронные сети в задачах комбинаторной оптимизации».

Литература

1. Drton M., Maathuis M.H. Structure Learning in Graphical Modeling // Annual Review of Statistics and Its Application. 2017. Vol. 4. P. 365–393. DOI: 10.1146/annurev-statistics-060116-053803.
2. Cordoba I., Bielza C., Larranaga P. A review of Gaussian Markov models for conditional independence // Journal of Statistical Planning and Inference. 2020. Vol. 206. P. 127–144. DOI: 10.1016/j.jspi.2019.09.008.
3. Kalyagin V., Kostylev I. Graph Density and Uncertainty of Graphical Model Selection Algorithms // Advances in Optimization and Applications. OPTIMA 2023. Vol. 1913 / eds. by N. Olenev, Y. Evtushenko, M. Jaćimović, M. Khachay, V. Malkova. Springer, Cham, 2023. P. 188–201. Communications in Computer and Information Science. DOI: 10.1007/978-3-031-48751-4_14.
4. Kalyagin V.A., Koldanov A.P., Koldanov P., Pardalos P.M. Statistical Analysis of Graph Structures in Random Variable Networks. Springer, 2020. DOI: 10.1007/978-3-030-60293-2.
5. Peng J., Wang P., Zhou N., Zhu J. Partial Correlation Estimation by Joint Sparse Regression Models // Journal of the American Statistical Association. 2009. Vol. 104. P. 735–746 DOI: 10.1198/jasa.2009.0126.