0.1. *Мещеряков Г.А.* Стохастический след и квадратуры Гаусса для учёта сложной структуры популяции

Линейные модели, и, в частности, структурные (англ. *structural equation models* или SEM), находят широкое применение в самых разнообразных областях — от социологии и эконометрики до биоинформатики и генетики [1]. Однако, часто такие модели используются в предположении о независимости наблюдений, которое на практике редко соблюдается: наблюдения могут иметь общую дисперсию, обусловленную некоторыми общими факторами. Например, в задаче полногеномного поиска ассоциаций (т.е. выявления значимых генов, влияющих на фенотипический признак), такими факторами могут быть генетическое родство наблюдаемых индивидов и/или особенности окружающей среды. Возможно учитывать такие зависимости путём выноса общих дисперсий в отдельные случайные величины, так называемые «случайные эффекты» (англ. *random effects*). В общем случае, линейная модель со случайными эффектами имеет вид:

$$\mathbf{Y} = \mathbf{A}(\theta) + \mathbf{E} + \sum_{k=1}^{R} \mathbf{U^{(k)}},$$
$$\mathbf{E} \sim MN(0, \Sigma, I_n), \mathbf{U^{(k)}} \sim$$
$$\sim MN(0, \mathbf{D^{(k)}}(\theta), \mathbf{K^{(k)}}(\theta)), \quad (1)$$

где Y — матрица наблюдений размера $n \times m$, A — матрица прямых эффектов размера $n \times m$, E — матрица случайных ошибок, $U^{(k)}$ — матрица случайных эффектов, $MN(M, B, C)$ — матричное нормальное распределение со средним M, матрицей ковариаций по столбцам B и по строкам C, n — число наблюдений, m — число наблюдаемых переменных, R — число случайных эффектов, θ — вектор параметров модели, подлежащий оцениванию.

При попытке получения оценки θ методом максимального правдоподобия для 1 мы неизбежно сталкиваемся с необходимостью нахождения обратной матрицы размера $n \times n$ при вычислении функции цели, что в общем случае возможно лишь за $O(n^3)$. В частном случае при $R = 1$ совместная диагонализация I_n и $K^{(1)}$ позволяет понизить сложность до $O(n)$ [2], однако при $R > 1$ это не представляется возможным. Было решено использовать подход на основе представления следа от аналитической матричной функции через интеграл Римана — Стильтеса и последующем приближением его через ортогональные полиномы Ланшоца [3]. Он основывается на следующих предпосылках:

1. Стохастическая оценка следа $f(A)$, где A — некоторая эрмитова матрица, f — аналитическая функция:
$$\forall x : E[x] = 0, E[x^T x] = 1 : E[x^T f(A) x] =$$
$$= tr\{f(A)\} \approx \frac{1}{N} \sum_{i=1}^{N} x_i^T f(A) x_i$$
2. Представление квадратичной формы в виде интеграла Римана — Стилтьеса и приближение его квадратурами Гаусса-Ланшоца:
$$x^T f(A) x = x^T f(Q^T \Lambda Q) x = x^T Q^T f(\lambda) Q x =$$
$$= \sum_{i=1}^{n} f(\lambda_i) \mu_i^2 = \int_{\lambda_n}^{\lambda_1} f(t) d\mu(t) \approx \sum_{i=1}^{m} w_i f(b_i)$$
$$\mu_i = (Qx)_i$$

Семейство ортогональных полиномов для квадратуры по мере μ находится через процесс Грамма — Шмидта, совпадающий в данной задаче с алгоритмом Ланшоца.

В рамках работы был реализован описанный функционал, позволяющий сократить сложность от числа элементов выборки до $O(n^2)$, и интегрирован в программный пакет **semopy**.

Список литературы

[1] Igolkina A., Meshcheryakov G., Gretsova M. et al. Multi-trait multi-locus SEM model discriminates SNPs of different effects // BMC genomics. 2020. Vol. 21 P. 1–11.

[2] Мещеряков Г. А., Иголкина А. А., Самсонова М. Г. Создание программного пакета для моделирования структурными уравнениями // Неделя науки СПбПУ. Санкт-Петербург: СПбПУ, 2018. С. 241–244.

[3] Ubaru S., Chen J., Saad Y. Fast Estimation of $tr(f(A))$ via Stochastic Lanczos Quadrature //SIAM Journal on Matrix Analysis and Applications. 2020. Vol. 38 (4). P. 1075–1099