Филиал Московского государственного университета
имени М.В. Ломоносова в г. Ташкенте

Гуломов Саидхужа Ботир угли

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

НА ТЕМУ: "Оценка вероятности попадания многомерной нормально распределенной случайной величины в первый октант"

ПО НАПРАВЛЕНИЮ 010500 - "Прикладная математика и информатика"

ВКР рассмотрена
и рекомендована к защите
зав. кафедрой "МаТИС"
д.ф.-м.н.,профессор
_____ Кудрявцев В.Б.

Научный руководитель
к.ф.-м.н., с.н.с.
_____ Алексеев Д.В.

Ташкент 2014

Содержание

Аннотация

При нормальном распределении часто возникает задача определения вероятности того, что случайный вектор попадает в первый октант. Например при нахождении вероятности ошибки декодирования при передаче кодового слова по Гауссовскому каналу. В работе рассматривается частный случай $n = 2$. Представлено решение задачи с помощью модели эллипсов рассеяния. Выведены формулы расчета вероятности посредством вычисления суммы площадей участков эллипсов. Проведена оценка погрешности при выборе данной модели.

Abstract

In a situation involving a gaussian distribution it happens to be a problem to calculate the probability that a random vector would fall onto the positive octant. For instance such problem exists when one tries to calculate the probability of a decoding error during the transmission of a code-word through the Gaussian channel. This paper considers an instance of 2-dimensional gaussian distribution. A solution is presented based on using the dispersion ellipses model. Formulas for calculation of the probabiltity by means of calculation of ellipse areas are carried out. A calculation error for the chosen model is estimated.

1 Введение

Задача вычисления вероятности попадания вектора в положительный октант сформировалась при дешифровке сигнала, проходящего по гауссовскому каналу, рекурсивным применением дешифрующей функции. Т.е. сигнал, состоящий из исходного сигнала и шума, при поступлении на дешифратор, многократно в нём обрабатывается. В результате, составляющие исходного сигнала, изначально независимые, становятся сильно зависимыми друг от друга. Матрица ковариаций при этом становится плохо обусловленной, а компоненты вектора средних - очень малыми.

В работах [3], [4] и [5] проведены исследования в проблеме вычисления вероятности нормально распределенной величины. Также созданы алгоритмы, вычисляющий с наперёд заданной точностью величину вероятности.

Однако особенностью подхода, рассмотренного в данной работе, является именно начальные параметры вероятности - плохо обусловленная ковариационная матрица, собственные значения которой близки к нулю, и вектор средних с очень малыми компонентами.

Подобного рода задача неправильно решается общеизвестными прикладными средствами, такими как пакет MatLAB.

В качестве закона распределения было выбрано нормальное в силу того, что оно является самой распространенной вероятностной моделью в мире.

В работе сначала представляется использование модели эллипсов рассеяния для подсчета вероятности. После показана реализация вычисления с помощью бесконечной суммы площадей слоёв эллипсоидов; проведено приближение до конечной суммы с последующей оценкой погрешности приближения.

Выражаю большую благодарность своему научному руководителю, Алексееву Дмитрию Владимировичу, за всевозможную поддержку, внимание и отзывчивость в процессе написания данной работы.

2 Основные определения и условные обозначения

2.1 Условные обозначения

Аналитическое представление эллипсоида через матрицу ковариаций и вектор средних

$$\ell(\vec{x}) = (\vec{x} - \vec{\mu})K^{-1}(\vec{x} - \vec{\mu})^T$$

Площадь области $i-$го эллипсоида, ограниченного осями координат и частью графика эллипсоида

$$S_i = \{(x, y) \in R^+ | \ell(x, y) \leq R_i\}$$

Площадь слоя эллипсоида - области между $i-$ым и $(i+1)-$ым эллипсоидами

$$D_i = S_{i+1} \backslash S_i$$

Функция вычисления площади слоя эллипсоида

$$s_i = S(D_i)$$

Произвольная точка на слое эллипсоида

$$\xi_i = \{(x, y)|(x, y) \in D_i\}$$

$$\ell(\xi_i) = \zeta_i$$

Функция плотности распределения

$$f(x) = \frac{1}{\sqrt{4\pi^2 detK}} e^{-\frac{1}{2}x}$$

2.2 Вспомогательные определения и теоремы

Преобразование Абеля

$$\sum_{k=1}^{N} a_k b_k = a_N b_N - \sum_{k=1}^{N-1} B_k (a_{k+1} - a_k),$$

где $B_k = \sum_{i=0}^{k} b_i$

Определение: r_n-тым остатком ряда $\sum_{k=1}^{\infty} a_k$ является ряд $\sum_{k=n+1}^{\infty} a_k$

Теорема(о среднем значении): Пусть $f(x)$ интегрируема в $[a;b]$ и пусть во всем этом промежутке $m \leq f(x) \leq M$; тогда

$$\int_a^b f(x)dx = \mu(b - a)$$

где $m \leq \mu \leq M$

Теорема(интегральный признак Коши-МакЛорена): пусть ряд $\sum_{n=1}^{\infty} a_n$ имеет форму

$$\sum_{n=1}^{\infty} a_n \equiv \sum_{n=1}^{\infty} f(n),$$

где $f(n)$ есть значение при $x = n$ некоторой функции $f(x)$, определенной для $x \geq 1$, непрерывной, положительной и монотонной. Тогда ряд $\sum_{n=1}^{\infty} f(n)$ сходится или расходится в зависимости от того, имеет ли функция

$$F(x) = \int f(x)dx$$

при $x \to +\infty$ конечный предел или нет.

2.3 Постановка задачи

Имеется 2–мерное нормальное распределение с вектором средних $\mu = (\mu_1, \mu_2)$ и матрицей ковариаций K, при этом заданы следующие условия на вектор средних и матрицу ковариаций:
$\mu = (\mu_1, \mu_2 \mid \mu_1 \leq 0, \mu_2 \leq 0)$ и определитель матрицы ковариаций равен 1, т.е. переменные x_1, x_2 линейно зависимы.
Плотность определяется формулой

$$p(x) = \frac{1}{\sqrt{(2\pi)^2 \cdot \det(K)}} \cdot \exp(-\frac{1}{2}(x-\mu) \cdot K^{-1} \cdot (x-\mu)^\top).$$

Необходимо найти вероятность попадания случайного вектора в первый октант, т.е.

$$\mathbb{R}^2_+ = \{(x_1, x_2) | x_1 \geq 0, x_2 \geq 0\}.$$

Указанная вероятность равна

$$P = \iint\limits_{x \in \mathbb{R}^2_+} p(x)dx.$$

Задача сводится к тому, чтобы вычислить данный интеграл с точностью ϵ

3 Основные результаты

Лемма 1. *Пусть даны:(1)единичная окружность с центром в точке O;(2)две хорды, пересекающиеся в точке K и (3) координаты начал и концов хорд - A*$(\cos\alpha, \sin\alpha)$*, B*$(\cos\beta, \sin\beta)$*, C*$(\cos\gamma, \sin\gamma)$*, D*$(\cos\delta, \sin\delta)$*(см. рис. 1).*

Тогда площадь фигуры CKB по формуле:

$$S_{BCK} = S_{\triangle BCK} + S_{BC}$$

Доказательство
Введём обозначения:

$$A = \sin\alpha - \sin\gamma$$

$$B = \cos\alpha - \sin\gamma$$

$$C = \sin\delta - \sin\beta$$

$$D = \cos\delta - \cos\beta$$

$$E = \sin\beta - \sin\gamma$$

$$F = \cos\beta - \cos\gamma$$

$$M = \frac{(EDB + AD\cos\gamma - CB\cos\beta)}{(AD - CB)}$$

На единичной окружности с центром в начале координат выбраны точки A, B, C, D.

Найти площадь сегмента, ограниченного окружностью и хордами AC и BD.

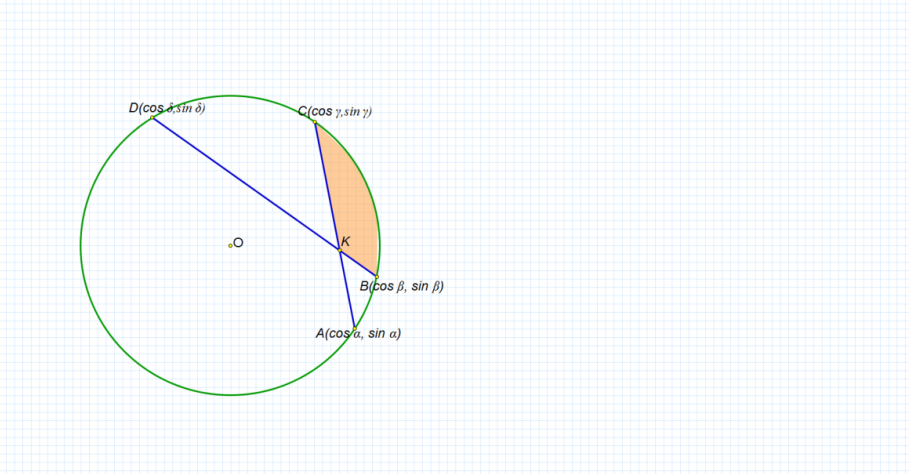

Рис. 1

$$p = \frac{1}{2}(\sqrt{F^2 - E^2} + |\cos\beta + \cos\gamma - 2M|\sqrt{(1 + (\frac{C}{D})^2)})$$

$$p_{\triangle BCO} = \frac{1}{2}(2 + \sqrt{F^2 - E^2})$$

$$S_{\triangle BCO} = \sqrt{p_{\triangle BCO}(p_{\triangle BCO} - 1)^2(p_{\triangle BCO} - \sqrt{F^2 - E^2})}$$

Таким образом:

$$S_{\triangle BCK} = \sqrt{p(p - \sqrt{F^2 - E^2})(p - |\cos\beta - M|\sqrt{(1 + (\frac{C}{D})^2)})(p - |\cos\gamma - M|\sqrt{(1 + (\frac{C}{D})^2)})}$$

$$S_{BC} = R^2 \arcsin(\frac{c}{2R}) - \frac{c}{4}\sqrt{4R^2 - c^2},$$

где R– радиус окружности.

Теорема 2. *Пусть дан бесконечный положительный ряд* $P = \sum_{i=1}^{\infty} s(D_i) \cdot f(\ell(\xi_i))$. *Тогда* $\forall \epsilon_1 > 0\ \exists N(\epsilon_1)$, *что при* $i > N$

$$\Big| \sum_{i=N+1}^{\infty} s(D_i) \cdot f(\ell(\xi_i)) \Big| < \epsilon_1$$

Доказательство
См. раздел 6.1

Теорема 3. *Пусть даны два конечных положительных ряда P_N и Q_N такие, что:*

$$P_N = \sum_{i=1}^{N} s(D_i) \cdot f(\zeta_i),\ Q_N = \sum_{i=1}^{N} s(D_i) \cdot f(R_i)$$

где $\zeta_i \in D_i$ - произвольное

Тогда существует такое $\epsilon_2 > 0$, что выполнено следующее:

$$|P_N - Q_N| < \epsilon_2$$

Доказательство
См. раздел 6.2

Следствие 4. *Общая погрешность ϵ алгоритма решения поставленной задачи равна:*

$$\epsilon = \epsilon_1 + \epsilon_2$$

4 Эллипсы рассеяния

Определение: Общее уравнение семейства эллипсов будет иметь вид

$$\frac{(x-\mu_x)^2}{\sigma_1^2} - \frac{2 \cdot \rho \cdot x \cdot y}{\sigma_1 \cdot \sigma_2} + \frac{(y-\mu_y)^2}{\sigma_2^2} = const$$

В рассматриваемом двумерном случае выберем в качестве геометрической модели рассмотрим так называемые эллипсы рассеяния, которые получаются при проецировании на плоскость xOy сечений кривой Гаусса, параллельных оси Ox По условию, матрица ковариаций имеет вид:

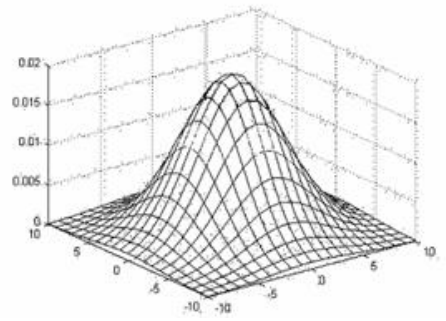

Рис. 2 Гауссиан

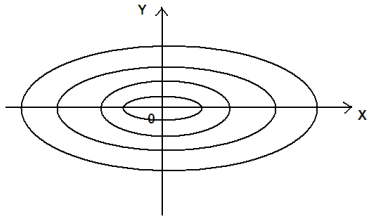

Рис.3 Эллипсы рассеяния

$$K = \begin{pmatrix} \sigma_1^2 & \rho \cdot \sigma_1 \cdot \sigma_2 \\ \rho \cdot \sigma_1 \cdot \sigma_2 & \sigma_2^2 \end{pmatrix},$$

где σ_1 и σ_2 - дисперсии, а ρ - коэффициент корреляции. В общем случае, главные оси симметрии семейства эллисов образуют с осью Ox угол $\alpha_i, i = \{1, 2\}$, тангенс которого определяется по следующей формуле:

$$\operatorname{tg} 2\alpha_i = \frac{2 \cdot \rho \cdot \sigma_1 \cdot \sigma_2}{\sigma_1^2 - \sigma_2^2}$$

Вычисление площади участка, ограниченного осями координат и графиком очередного эллипса - задача трудоёмкая, поэтому воспользуемся оператором аффинного преобразования - оператором сжатия, применим его к всему семейству эллипсов и в результате получим семейство окружностей.

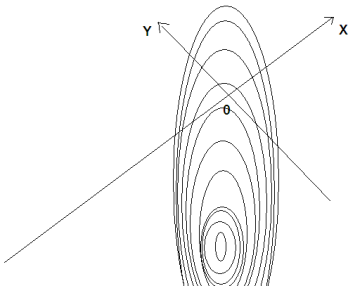

Рис.4 Эллипсы рассеяние исходной задачи

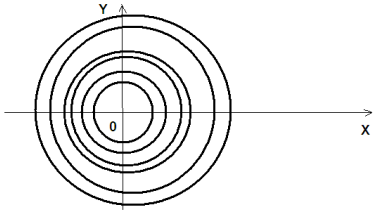

Рис. 5 Эллипсы рассеяния после смещения, поворота и сжатия(семейство окружностей)

Для того, чтобы найти площадь участка, ограниченного осями координат и графиком окружности, решим следующую общую задачу, представленную на рис. 1.

Площадь необходимой нам фигуры, заключенной между дугой $\breve{BC}$ и 2-мя пересекающимися хордами, есть сумма площади треугольника $\triangle BCK(S_{\triangle BCK})$, и кругового сегмента хорды BC и дуги $\breve{BC}(S_{BC})$.

5 Аппроксимация интеграла с помощью бесконечной суммы

Как указано в разделе "Постановка задачи", полная вероятность попадания произвольной точки равна:

$$P = \iint\limits_{x \in R_+^n} p(x, y) dx dy.$$

Согласно избранному методу решения задачи с помощью семейства эллипсов рассеяния, полная вероятность принимает вид:

$$P = \sum_{i=1}^{\infty} \int\limits_{D_i} p(x, y)\ dx dy,$$

где $D_i = \{(x,y) \in R^+ | R_i < l(x,y) \leq R_{i+1}\}$.

Так как функция $p(x,y)$ интегрируема в области D_i, и выполнено неравество $\frac{1}{\sqrt{4\pi^2 detK}} \exp\{-\frac{1}{2}R_{i+1}\} \leq p(x,y) \leq \frac{1}{\sqrt{4\pi^2 detK}} \exp\{-\frac{1}{2}R_i\}$, воспользуемся теоремой о среднем и получим:

$$P = \sum_{i=1}^{\infty} s(D_i) \cdot f(\ell(\xi_i)),$$

Пусть $\ell(\xi_i) = \zeta_i$. Получаем:

$$p(x) = \frac{1}{\sqrt{4\pi^2 detK}} e^{-\frac{1}{2}(x-\mu)K^{-1}(x-\mu)^T} \Rightarrow f(\zeta_i) = \frac{1}{\sqrt{4\pi^2 detK}} e^{-\frac{1}{2}\zeta_i}$$

Производная от $f(\zeta_i)$ равна:

$$f'(\zeta_i) = -\frac{1}{2\sqrt{4\pi^2 detK}} e^{-\frac{1}{2}\zeta_i}$$

Зафиксируем на области D_i точку $\tilde{\zeta}$ и рассмотрим сумму вида

$$P_N = \sum_{i=1}^{N} s(D_i) \cdot f(\tilde{\zeta}),$$

где $N = N(\epsilon)$, ϵ - верхняя оценка остатка ряда P в процессе аппроксимации.

Данный ряд показывает возможность реализации подсчёта вероятности по данным задачи на ЭВМ. Для организации конечного времени вычислений, необходимо задавать порог, при котором значение вероятности P бесконечно мало, т.е. задать точность ϵ.

В данном случае искомая величина ϵ будет суммой двух видов погрешности: ϵ_1(погрешность аппроксимации бесконечной суммы ряда, задаваемая как верхняя оценка остатка этого ряда), и ϵ_2(погрешность при фиксировании точки на области D_i)

6 Оценка погрешностей

6.1 Оценка сверху остатка бесконечного ряда P

Для оценки погрешности аппроксимации, воспользуемся преобразованием Абеля и применим его к приближенной формуле вычисления вероятности. Оценим r_n-тый остаток приближенной формулы вычисления вероятности. По формуле преобразования Абеля получаем:

$$\sum_{i=n+1}^{\infty} s(D_i) f(\zeta_i) \Rightarrow \lim_{A\to\infty} \left| \sum_{i=n+1}^{A} s(D_i) f(\zeta_i) \right|$$

$$\Rightarrow \lim_{A\to\infty} \left| (f(\zeta_A) s(D_A) - \sum_{i=n+1}^{A-1} \big(f(\zeta_{i+1}) - f(\zeta_i)\big) H_i) \right|,$$

где $H_i = \sum_{m=0}^{i} s(D_m)$ – положим ограниченная числовая последовательность, $|H_i| \leq L$, где $L = const,\ L > 0$

Исходя из вида функции $f(\zeta_A)$, при $\lim_{A\to\infty} f(\zeta_A) \to 0$, а $s(D_A)$ - величина конечная и, следовательно, ограниченная. Получаем:

$$\lim_{A\to\infty} \left| \sum_{i=n+1}^{A-1} \big(f(\zeta_{i+1}) - f(\zeta_i)\big) H_i \right|$$

Рассмотрим график функции $f(x),\ 0 \leq x \leq +\infty$

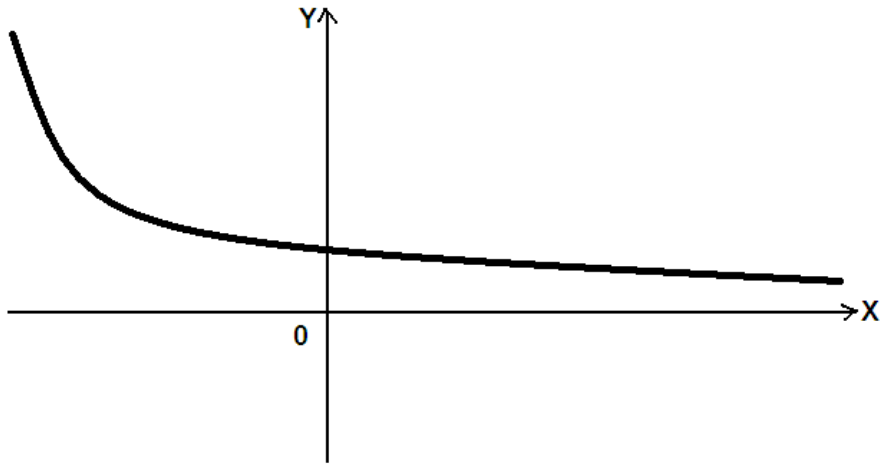

Рис. 6 График функции $f(x),\ 0 \leq x \leq +\infty$

Из рисунка видно, что при $x \to +\infty :\ f(x) \to 0$. Следовательно $\exists\tilde{\epsilon} > 0$, что $\sum_{i=n+1}^{\infty} |f(\zeta_{i+1}) - f(\zeta_i)| < \tilde{\epsilon}$. Т.к. $|H_i| < L$, то положим значение $\tilde{\epsilon} = \frac{\epsilon_1}{L}$. Получим:

$$\lim_{A\to\infty} \left| \sum_{i=n+1}^{A-1} \big(f(\zeta_{i+1}) - f(\zeta_i)\big) H_i \right| < \frac{\epsilon_1}{L} \cdot L = \epsilon_1$$

6.2 Оценка сверху погрешности вычисления интеграла на слое эллипсоида

При рассмотрении вида двумерного гауссиана, удовлетворяющего данным задачи, становится видно, что график гауссиана, со значениями абсциссы из первого квадранта, представляет собой бесконечно убывающую функцию. Эта функция обозначена как $f(x), 0 \leq x \leq +\infty$.

При выводе ряда аппроксимации P_N была предварительно зафиксирована произвольная точка на области D_i. Для получения верхней оценки

погрешности такой фиксации воспользуемся свойством бесконечного убывания графика функции $f(x)$.

Итак, зафиксируем точку ζ_i на области D_i

Т.к. выполнено $R_i \leq \zeta_i \leq R_{i+1}$, то $f(R_i) \geq f(\zeta_i) \geq f(R_{i+1})$. Исходя из намерения получить верхнюю оценку, то выберем в качестве $f(\zeta_i) = max\{f(R_i), f(R_{i+1})\}$. Получим

$$Q_N = \sum_{i=1}^{N} s(D_i) f(R_i)$$

Теперь, зная Q_N, задача состоит в оценке следующего выражения:

$$|P_N - Q_N| = \sum_{i=1}^{N} s(D_i)(f(R_i) - f(\tilde{\zeta}))$$

Т.к. D_i представляет собой конечную область, то и площадь этой области будет конечной, соответственно ряд $\sum_{i=1}^{N} s(D_i)$ является ограниченным некоторым наперед заданным числом $L > 0$. Таким образом:

$$|P_N - Q_N| < L|\sum_{i=1}^{N} f(R_i) - f(\tilde{\zeta})|$$

Обращая внимание на рис. 7, видно, что часть графика, попадающего в первый квадрант, представляет непрерывную бесконечно убывающую функцию. Следовательно, существует такое $\tilde{\epsilon} = \frac{\epsilon_2}{L} > 0$, что выполнено:

$$L|\sum_{i=1}^{N} f(R_i) - f(\tilde{\zeta})| \leq L \cdot \tilde{\epsilon} = L \cdot \frac{\epsilon_2}{L} = \epsilon_2$$

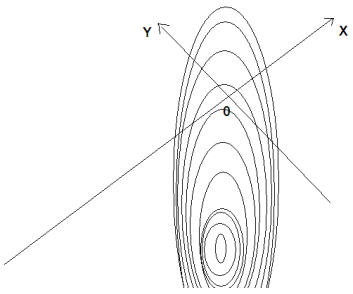

Рис.7 Разрез трехмерного гауссиана на эллипсы рассеяния

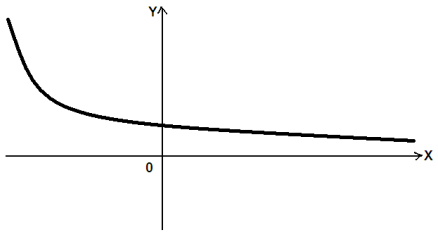

Рис. 8 Часть гауссиана, попадающая в первый октант

7 Выводы

Приведена математическая модель решения поставленной задачи - построение семейства эллипсов рассеяния. Получено решение поставленной задачи посредством вычисления площадей областей, расположенных между соседними эллипсов и их последующем суммировании. Доказана аппроксимация бесконечного интеграла P конечной суммой P_N. Проведена оценка погрешности данной аппроксимации. Решение задачи реализовано в виде MFC-приложения.

Список литературы

[1] Александров П.С. *Курс аналитической геометрии и линейной алгебры* - М.: Наука, Главная редакция физико-математической литературы, 1979

[2] Вентцель Е.С. *Теория вероятностей: Учеб. для вузов., 6-е изд. стер.* - М.: Высш. шк., 1999

[3] D.R.Cox, N. Wermuth *A Simple Approximation for Bivariate and Trivariate Normal Integrals* - International Statistical Review (1991), 59, 2, pp. 263-269

[4] Alan Genz *Numerical Computation of Multivariate Normal Probabilities* - Journal of Computational and Graphical Statistics, vol.1, No. 2, Jun., 1992

[5] Alan Genz, Koon-Sing Kwong *Numerical Evaluation of Singular Multivariate Normal Distributions* - revised version published in J. Stat. Comp. Simul. 68 (2000), pp. 1-21.