Множественная проверка гипотез на пространственных структурах. Применение теории случайных полей.

Рябенко Евгений, аспирант

10 декабря 2010 г.

1. Множественная проверка гипотез
   - Классическая схема проверки гипотезы
   - Проверка большого числа гипотез
   - FWER
2. Проверка гипотез на пространственных структурах
   - Примеры задач
   - Теория случайных полей
   - Перестановочные методы

Математическая формулировка

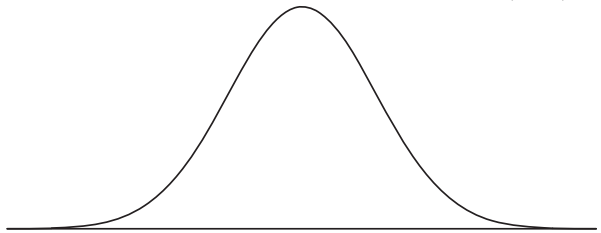
выборка: $\mathbf{X} = \{X_1, \ldots, X_n\} \sim P \in \Omega;$

нулевая гипотеза: $H_0 : P \in \omega, \ \ \omega \in \Omega;$

альтернатива: $H_1 : P \notin \omega;$

статистика: $T(\mathbf{X}), \ \ T(\mathbf{X}) \sim F(x)$ при $P \in \omega;$

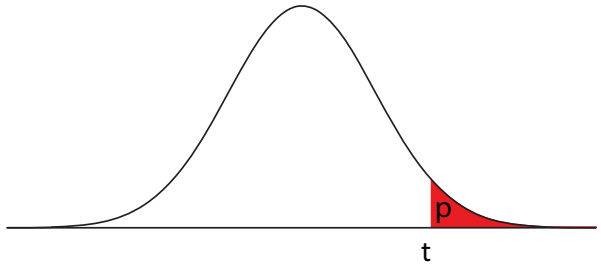
$T(\mathbf{X}) \not\sim F(x)$ при $P \notin \omega;$



реализация выборки: $\mathbf{x} = \{x_1, \ldots, x_n\};$

реализация статистики: $t = T(\mathbf{x});$

достигаемый уровень значимости: $p(\mathbf{x})$— вероятность при H_0 получить $T(\mathbf{X}) = t$ или ещё более экстремальное;



Гипотеза отвергается при $p(\mathbf{x}) \leq \alpha, \ \ \alpha$— уровень значимости.

Правило проверки гипотезы

Простейший пример

Требуется проверить симметричность монеты за 10 подбрасываний.

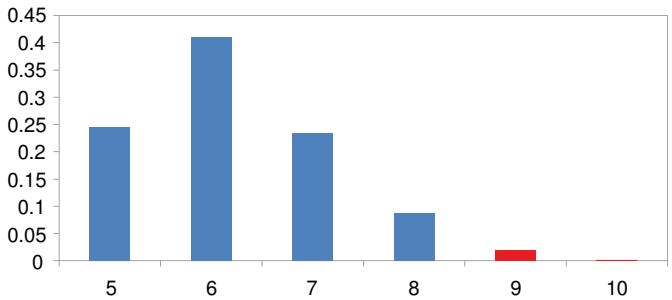
выборка: $\mathbf{X} = \{X_1, \ldots, X_{10}\} \sim P \in Bin\,(10, q)\,;$

нулевая гипотеза: $H_0 \colon P = Bin\,(10, 0.5)\,;$

альтернатива: $H_1 \colon P \neq Bin\,(10, 0.5)\,;$

статистика: $T\,(\mathbf{X}) = \max\,(\sum X_i, 10 - \sum X_i)\,;$

большие значения статистики свидетельствуют в пользу H_1;



реализация выборки: $\mathbf{x} = \{x_1, \ldots, x_{10}\};$

реализация статистики: $t = T\,(\mathbf{x})\,;$

достигаемый уровень значимости: при $t = 10 \;\; p = \frac{1}{512} \approx 0.002,$

при $t = 9 \;\; p = \frac{11}{512} \approx 0.02.$

При $t \geq 9$ гипотеза отвергается на уровне значимости $\alpha = 0.05$.

Несимметричность задачи проверки гипотез

	H_0 верна	H_0 неверна
H_0 принимается	H_0 верно принята	Ошибка второго рода
H_0 отвергается	Ошибка первого рода	H_0 верно отвергнута

Вероятность ошибки первого рода жёстко ограничивается достаточно малой наперёд заданной величиной — $P\left(p\left(\mathbf{x}\right) \leq \alpha \,|H_0\,\right) \leq \alpha$.

Вероятность ошибки второго рода минимизируется путём выбора достаточно мощного критерия.

Усложнение примера

Требуется проверить симметричность 1000 монет.

Пусть все монеты симметричны.
Вероятность того, что хотя бы одна 10 раз за серию упадёт одной и той же стороной, равна

$$1 - \left(1 - \frac{1}{512}\right)^{1000} \approx 0.86.$$

Вероятность того, что хотя бы одна не менее 9 раз за серию упадёт одной и той же стороной, равна

$$1 - \left(1 - \frac{11}{512}\right)^{1000} \approx 0.9999999996.$$

Следовательно, при проведении статистического анализа данных по большому количеству гипотез необходимо ограничивать не только вероятность каждой ошибки первого рода, но и некую глобальную меру ошибки, учитывающую число гипотез.

Математическая постановка

данные: $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\} \sim P \in \Omega;$

нулевые гипотезы: $H_i \colon P \in \omega_i, \quad \omega_i \in \Omega;$

альтернативы: $H_i' \colon P \notin \omega_i;$

статистики: $T_i = T(\mathbf{X}_i)$ проверяет H_i против $H_i';$

реализации статистики: $t_i = t(\mathbf{x}_i);$

достигаемые уровни значимости: $p_i = p(\mathbf{x}_i), \quad i = 1, \ldots, m;$

$\mathbf{M} = \{1, 2, \ldots, m\};$

$H_0 = \bigcup_{i \in \mathbf{M}} H_i$ — полная нулевая гипотеза;

$\mathbf{M_0} = \mathbf{M_0}(P) = \{i \colon H_i \text{ верна}\}$ — индексы верных гипотез, $|\mathbf{M_0}| = m_0;$

$\mathbf{R} = \mathbf{R}(P, \alpha) = \{i \colon H_i \text{ отвергнута}\}$ — индексы отвергаемых гипотез, $|\mathbf{R}| = R;$

$V = |\mathbf{M_0} \cap \mathbf{R}|$ — число ошибок первого рода.

	Число верных H_0	Число неверных H_0	Всего
Число принятых H_0	U	T	$m - R$
Число отвергнутых H_0	V	S	R
Всего	m_0	$m - m_0$	m

Многомерные обобщения числа ошибок первого рода

Групповая вероятность ошибки (первого рода):

$$FWER = P\left(V \geq 1\right).$$

Контроль над групповой вероятностью ошибки на уровне α означает

$$FWER = P\left(V \geq 1\right) \leq \alpha \quad \forall P.$$

В задачах проверки достаточно большого числа гипотез обеспечение контроля над $FWER$ может приводить к росту консервативности статистической процедуры.

Ожидаемая доля ложных отклонений гипотез (среди всех отклонений):

$$FDR = \mathbb{E}\left(\frac{V}{R} \cdot I\left(R > 0\right)\right).$$

Контроль над ожидаемой долей ложных отклонений на уровне q означает

$$FDR = \mathbb{E}\left(\frac{V}{R} \cdot I\left(R > 0\right)\right) \leq q \quad \forall P.$$

Контроль над групповой вероятностью ошибки

$\alpha^* \leq \alpha$ — уровень значимости, на котором необходимо проверять гипотезы $H_1, \ldots, H_m$; задача — выбрать его так, чтобы обеспечить $FWER \leq \alpha$;
t^* — соответствующее ему критическое значение статистики T.

Слабый контроль над $FWER$:

$$FWER = P(V \geq 1) = P\left(\bigcup_{i \in \mathbf{M}} \{p_i \leq \alpha^*\} \,|H_0\right)$$
$$= P\left(\bigcup_{i \in \mathbf{M}} \{T_i \geq t^*\} \,|H_0\right) \leq \alpha.$$

Сильный контроль над $FWER$:

$$FWER = P(V \geq 1) = P\left(\bigcup_{i \in \mathbf{M}^*} \{p_i \leq \alpha^*\} \,\middle|\, \bigcup_{i \in \mathbf{M}^*} H_i\right)$$
$$= P\left(\bigcup_{i \in \mathbf{M}^*} \{T_i \geq t^*\} \,\middle|\, \bigcup_{i \in \mathbf{M}^*} H_i\right) \leq \alpha \;\; \forall \mathbf{M}^* \subset \mathbf{M}.$$

Subset pivotality

Из слабого контроля следует сильный, если для задачи выполняется свойство **subset pivotality**: нулевое распределение любого подмножества статистик T_i не зависит от того, верны или неверны соответствующие оставшимся статистикам гипотезы.

$$P\left(\bigcup_{i\in\mathbf{M}^*}\{T_i \geq t^*\}\middle|\bigcup_{i\in\mathbf{M}^*}H_i\right) = P\left(\bigcup_{i\in\mathbf{M}^*}\{T_i \geq t^*\}|H_0\right)$$
$$\leq P\left(\bigcup_{i\in\mathbf{M}}\{T_i \geq t^*\}|H_0\right)$$

Максимальная статистика

$M_T = \max_i T_i$ — максимальная статистика.

$$\bigcup_i \{T_i \geq t^*\} = \{M_T \geq t^*\}.$$

Для обеспечения $FWER \leq \alpha$ достаточно знать распределение максимальной статистики при справедливости полной нулевой гипотезы $F_{M_T|H_0}(x)$:

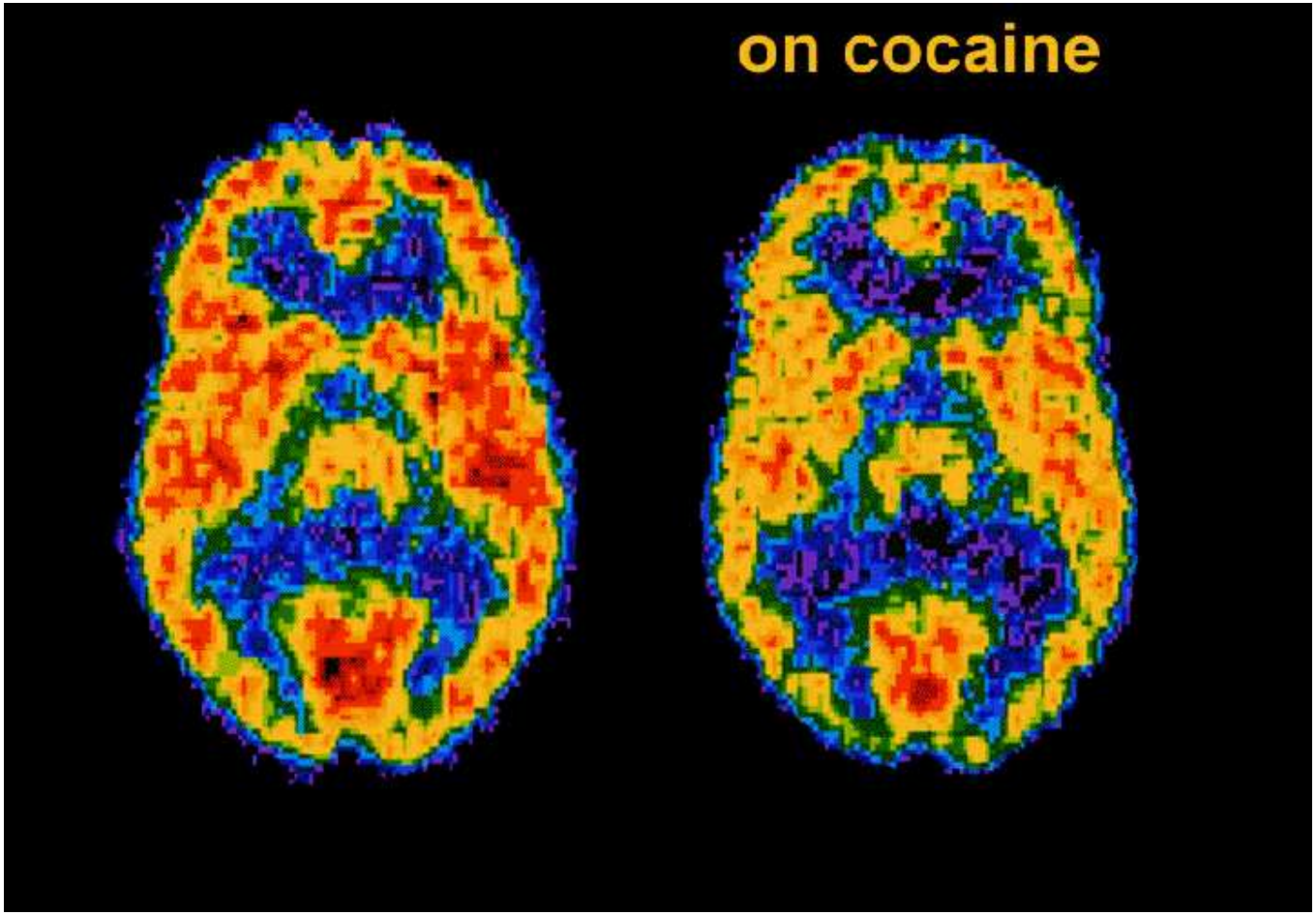
$$t_\alpha \equiv F^{-1}_{M_T|H_0}(1-\alpha),$$

$$\begin{aligned} P\left(\bigcup_{i\in\mathbf{M}} \{T_i \geq t^*\}\,|H_0\right) &= P\left(M_T \geq t_\alpha\,|H_0\right) \\ &= 1 - F_{M_T|H_0}(t_\alpha) \\ &= \alpha. \end{aligned}$$

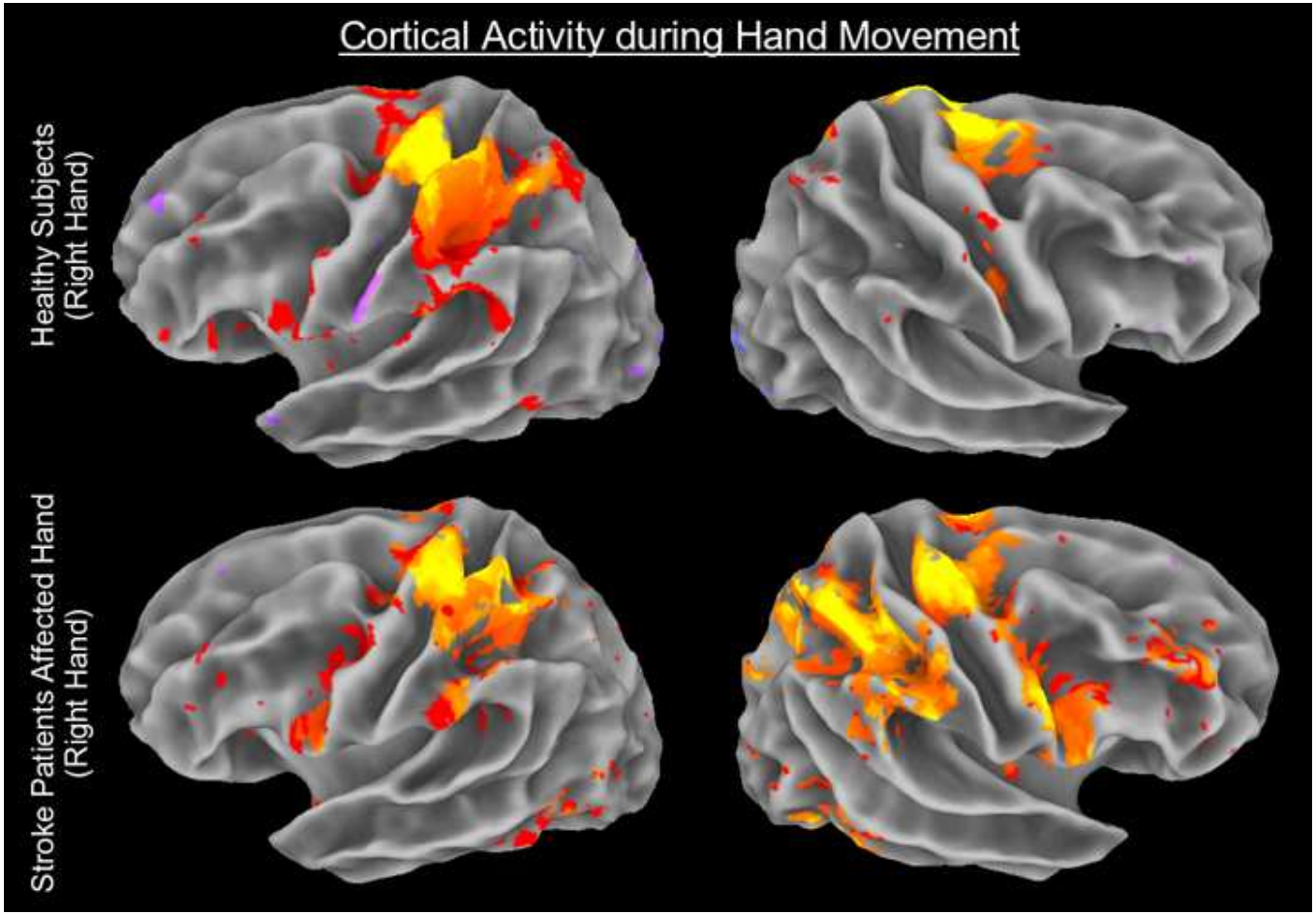
1. Множественная проверка гипотез
   - Классическая схема проверки гипотезы
   - Проверка большого числа гипотез
   - FWER

2. Проверка гипотез на пространственных структурах
   - Примеры задач
   - Теория случайных полей
   - Перестановочные методы

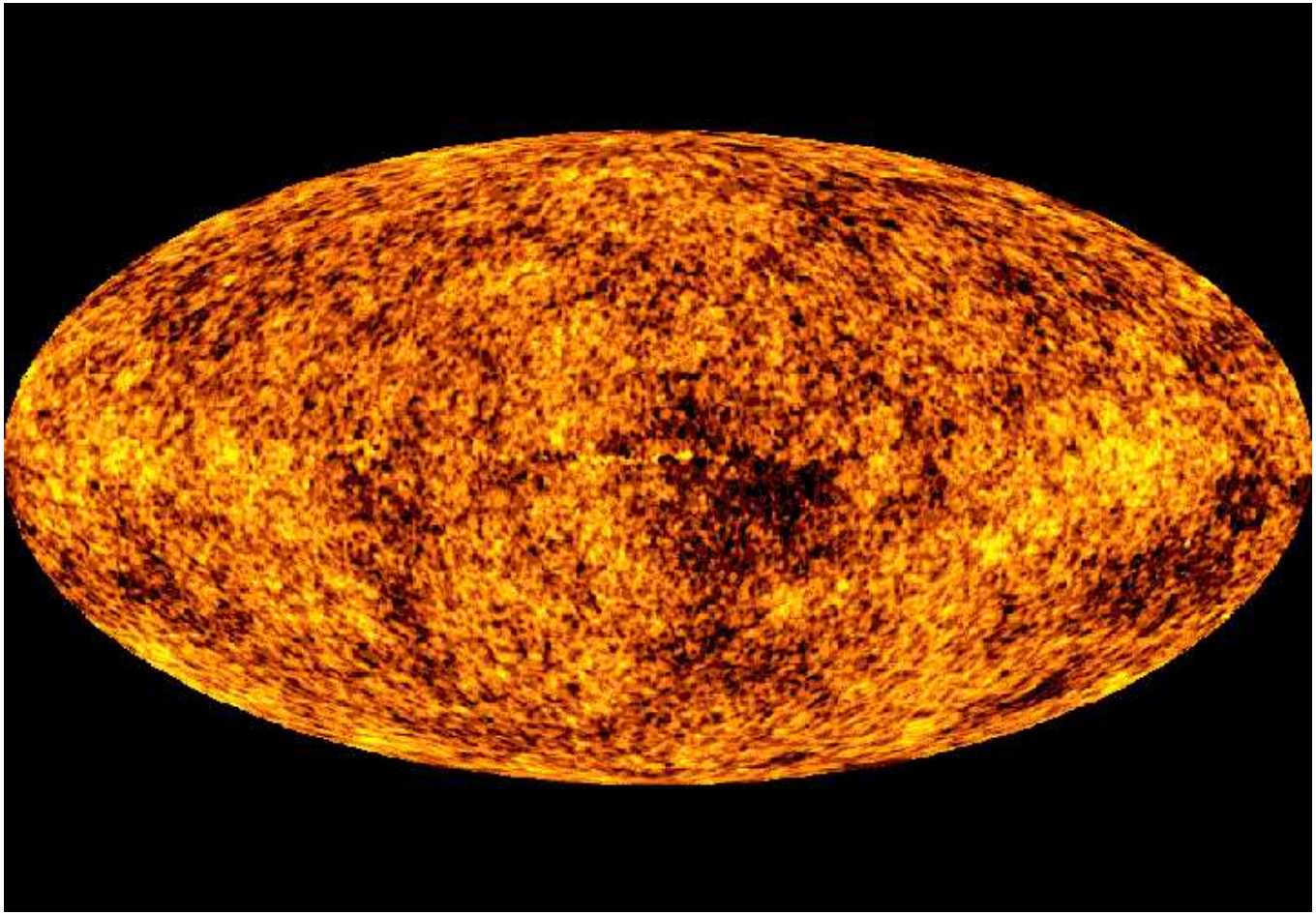
Позитронно-эмиссионная томография



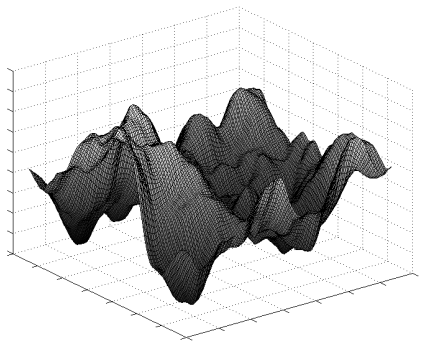
Магнитно-резонансная томография



Реликтовое излучение



Постановка задачи и обозначения



- $f(M)$ — случайное поле, зависящее от параметра $M \subset \mathbb{R}^N, \quad N \geq 1$
- $m(x) = E\{f(x)\}$ — среднее
- $R(x,y) = E\left\{(f(x) - m(x))\left(\overline{f}(y) - \overline{m}(y)\right)\right\}$ — ковариационная функция
- Если $m = const$ и $R(x,y) = R(x-y)$, то f **стационарно**
- Если при этом $R(x) = R(\|x\|)$, то то f **изотропно**
- Если случайные величины $f(x_1), \ldots, f(x_k)$ имеют многомерное гауссово распределение $\forall x_1, \ldots, x_k \in M, \quad \forall k \geq 1$, то случайное поле **гауссово**
- $A_u = A_u(f, M) = \{x \in M : f(x) \geq u\}$ — отклоняющееся множество.

Задача: аппроксимация вероятности отклонения

$$P\left\{\sup_{x \in M} f(x) \geq u\right\} = P\{A_u \neq \varnothing\} = P(M_T \geq t_\alpha \,|H_0) = FWER.$$

LKC

Lipschitz-Killing curvatures (Quermassintegrales, Minkowski functionals, Steiner functionals, integral curvatures, intrinsic volumes): множество характеристик $\mathfrak{L}_0(A), \ldots, \mathfrak{L}_N(A)$ N-мерного отклоняющегося множества A.

При $N = 2$:

- $\mathfrak{L}_2(A)$ — площадь A,
- $\mathfrak{L}_1(A)$ — длина границы A,
- $\mathfrak{L}_0(A)$ — Эйлерова характеристика A,
  $\mathfrak{L}_0(A) = \sharp\{\text{связные компоненты } A\} - \sharp\{\text{"дырки"в } A\}$.

При $N = 3$:

- $\mathfrak{L}_3(A)$ — объём A,
- $\mathfrak{L}_2(A)$ — половина площади поверхности A,
- $\mathfrak{L}_1(A)$ — удвоенный калибр (caliper diameter) A,
- $\mathfrak{L}_0(A)$ — Эйлерова характеристика A, $\mathfrak{L}_0(A) =$
  $\sharp\{\text{связные компоненты } A\} - \sharp\{\text{"ручки"в } A\} + \sharp\{\text{"дырки"в } A\}$.

$$\left| P\left\{ \sup_{x \in M} f(x) \geq u \right\} - E\left\{ \mathfrak{L}_0\left(A_u(f, M)\right) \right\} \right| \leq error(u).$$

Эйлерова характеристика

Определение.

$$\phi(A) = \begin{cases} 0, A = \varnothing, \\ 1, A \text{ топологически эквивалентно шару,} \end{cases}$$

$$\phi(A \cup B) = \phi(A) + \phi(B) - \phi(A \cap B).$$

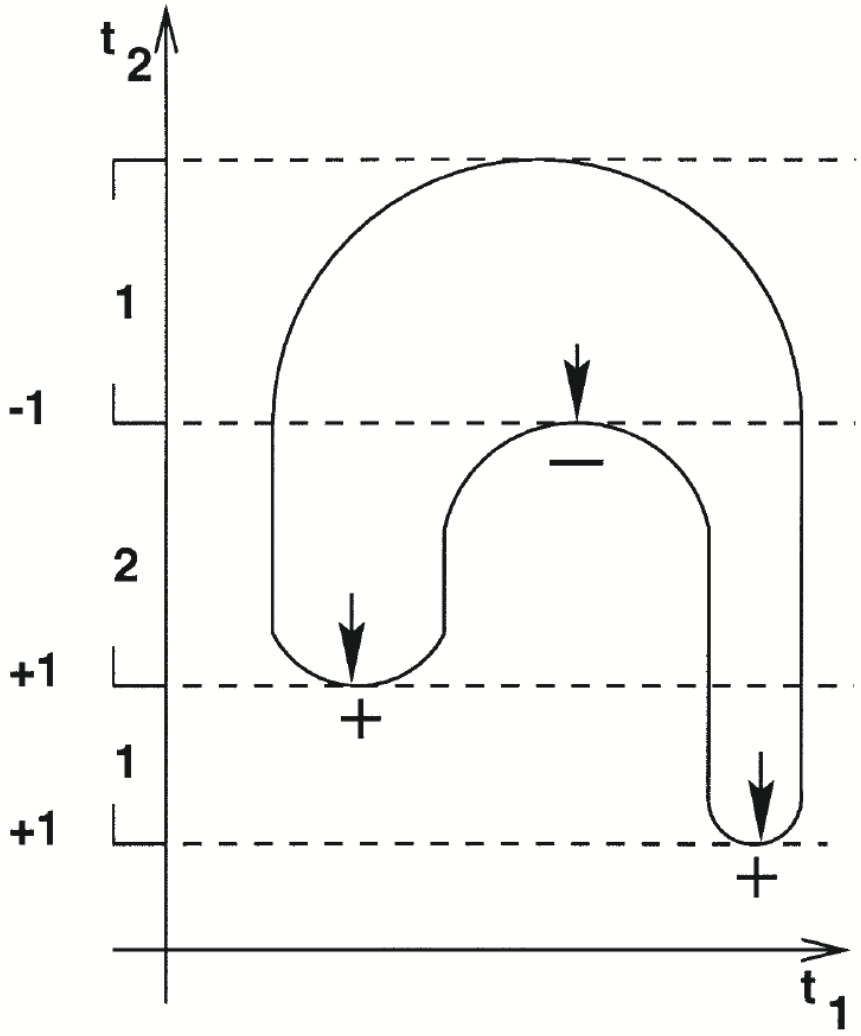
Итеративное определение, верно для достаточно "хороших" A.

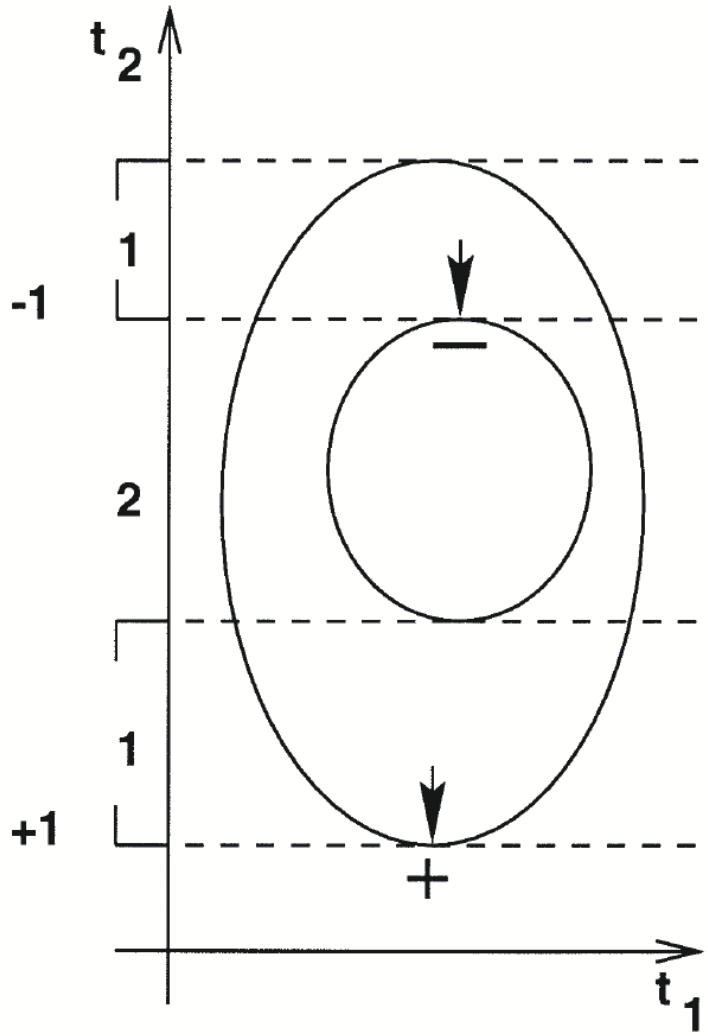
$$\phi(A) = \begin{cases} \text{Число непересекающихся интервалов в } A, N = 1, \\ \sum \{\phi(A \cap \varepsilon_x) - \phi(A \cap \varepsilon_{x-})\}, N > 1, \end{cases}$$

$$\phi(A \cap \varepsilon_{x-}) = \lim_{y \downarrow 0} \phi(A \cap \varepsilon_{x-y}),$$

ε_x — $(N-1)$-мерная плоскость в $\mathbb{R}^N$, состоящая из точек, j-я координата которых равна x.

Эйлерова характеристика





Требования к M

Наиболее общее ограничение: M — **расслоённое многообразие Уитни**.

- Конечные базовые комплексы: N-мерные прямоугольники вида $T = \prod_{i=1}^{M} [0, T_i]$, их конечные объединения.
- Римановы многообразия (N-мерные шары, сферы, их гладкие деформации).
- Множества, которые могут быть записаны в форме $M = \bigcup_{i=1}^{N} \partial_i M$, где ∂_i — i-мерные множества, относящиеся к одной из предыдущих категорий.

Пример: fMRI, значение в каждом вокселе — среднее по малой эллипсоидальной окрестности вокселя. Эллипсоид задаётся тремя осями, вдоль которых выбираются его размеры, и двумя углами поворота. Итог: скан fMRI — реализация случайного поля в восьмимерном пространстве

$$\{\text{координаты вокселя}\} \times \{\text{размеры эллипса}\} \times \{\text{углы поворота эллипса}\}.$$

Требования к f

- $f\colon M \in \mathbb{R}^N \to \mathbb{R}^k$ — гладкое гауссово случайное поле
- $m(x) = 0$ (для удобства, главное, чтобы $m(x) = const$)
- f имеет постоянную дисперсию на M (ослабление требования стационарности)
- все k компонент f дважды непрерывно дифференцируемы

$$f(x) \to f_{standartized}(x) = \frac{f(x)}{\left(E\left(f^2(x)\right)\right)^{1/2}}.$$

Окологауссовы случайные поля

Случайное поле $f\colon M \to \mathbb{R}^d$ назовём окологауссовым, если можно найти поле

$$g(x) = (g_1(x), \ldots, g_k(x)) : M \to \mathbb{R}^k$$

с независимыми одинаково распределёнными компонентами с нулевым средним и единичной дисперсией, и такую функцию

$$F\colon \mathbb{R}^k \to \mathbb{R}^d,$$

что f имеет то же многомерное распределение, что и $F(g)$.

- $F = \sum_1^k x_i^2$ — χ^2-поле с k степенями свободы
- $F = \frac{x_1\sqrt{k-1}}{\left(\sum_2^k x_i^2\right)^{1/2}}$ — T-поле с $k-1$ степенями свободы
- $F = \frac{m\sum_1^n x_i^2}{n\sum_{n+1}^{n+m} x_i^2}$, $k = m+n$, — F-поле с n и m степенями свободы

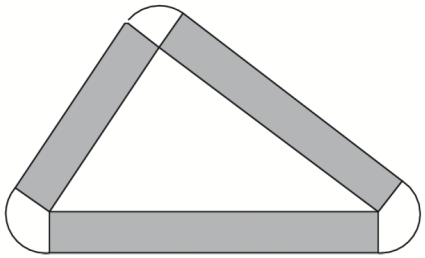
$$A_{[u,\infty)}(f, M) = A_{[u,\infty)}(F(g), M) = \{x \in M\colon (F \circ g)(x) \in [u,\infty)\}$$
$$= \left\{x \in M\colon g(x) \geq F^{-1}([u,\infty))\right\} = A_{F^{-1}([u,\infty))}(g, M).$$

Дополнительное требование: F — дважды непрерывно дифференцируема.

Tube formulae

Расширением (enlargement, tube) множества $A \in \mathbb{R}^N$ диаметра ρ назовём множество

$$Tube\left(A,\rho\right)=\left\{x\in\mathbb{R}^N:\ \min_{y\in A}\|x-y\|\leq\rho\right\}.$$



Теорема Стейнера (Steiner): если λ_N — объём в $\mathbb{R}^N$, то для выпуклого A

$$\lambda_N\left(Tube\left(A,\rho\right)\right)=\sum_{j=0}^{dim(A)}\omega_{N-j}\rho^{N-j}\mathfrak{L}_j\left(A\right),$$

где $\omega_j=\pi^{j/2}/\Gamma\left(1+j/2\right)$ — объём единичного шара в $\mathbb{R}^j$.
При достаточно малом ρ все рассматриваемые здесь множества являются выпуклыми.

GMF

Нас интересует содержание вероятностной меры в множестве $D = [u, \infty)$.
Пусть X — вектор k i.i.d. стандартных гауссовых случайных величин, тогда запишем

$$\gamma_k(A) = P\{X \in A\} = \int_A \frac{e^{-\|x\|^2/2}}{(2\pi)^{k/2}} dx.$$

Для расширения A для достаточно малого ρ справедливо разложение

$$\gamma_k(Tube(A, \rho)) = \sum_{j=0}^{\infty} \frac{\rho^j}{j!} \mathcal{M}_j^k(A),$$

где $\mathcal{M}_j^k(A)$ — гауссов функционал Минковского.
Например, для $k = 1$ и $A = [u, \infty)$

$$\mathcal{M}_j^k([u, \infty)) = H_{j-1}(u) \frac{e^{-u^2/2}}{\sqrt{2\pi}},$$

$$H_n(x) = n! \sum_{j=0}^{\lfloor n/2 \rfloor} \frac{(-1)^j x^{n-2j}}{j!(n-2j)!2^j}, \quad n \geq 0,$$

$$H_{-1}(x) = \sqrt{2\pi}\Psi(x) e^{x^2/2}.$$

Пример GMF для окологауссова поля

Для χ^2-поля с k степенями свободы для $j \geq 1$

$$\mathcal{M}_j^k\left(F^{-1}\left([u,\infty)\right)\right) = \frac{u^{k-j}e^{-u^2/2}}{\Gamma\left(k/2\right)2^{(k-2)/2}} \sum_{l=0}^{\lfloor \frac{j-1}{2} \rfloor} \sum_{m=0}^{j-1-2l} 1_{\{k \geq j-m-2l\}} C_{k-1}^{j-1-m-2l}$$
$$\times \frac{(-1)^{j-1+m+l}\left(j-1\right)!}{m!l!2^l} u^{2m+2l}.$$

Основной результат

$f\colon M \to \mathbb{R}^k$ — изотропное гауссово случайное поле с i.i.d. компонентами, нулевым средним, постоянной единичной дисперсией и вторым спектральным моментом

$$\lambda_2 = \left.\frac{\partial^2 R(x)}{\partial x_i^2}\right|_{x=0}.$$

Тогда для множества M размерности N и множества D размерности k справедливо

$$E\left\{\mathfrak{L}_i\left(A_D\left(f, M\right)\right)\right\} = \sum_{j=0}^{N-i} \begin{bmatrix} i+j \\ j \end{bmatrix} \frac{\lambda_2^{i+j}}{(2\pi)^{j/2}} \mathfrak{L}_{i+j}\left(M\right) \mathcal{M}_j^k\left(D\right),$$

где

$$\begin{bmatrix} n \\ j \end{bmatrix} = C_n^j \frac{\omega_n}{\omega_{n-j}\omega_j}.$$

{Выражение для $E\left\{\mathfrak{L}_0\left(A_u\left(f, T\right)\right)\right\}$ на N-мерном прямоугольнике}

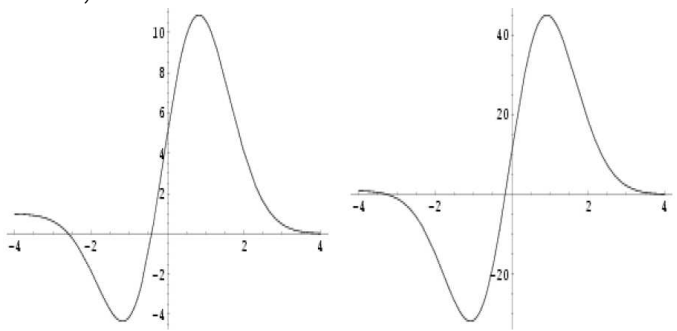
{Выражение для $E\left\{\mathfrak{L}_0\left(A_u\left(f, T\right)\right)\right\}$ на N-мерном кубе}

$N = 2, \quad \sigma^2 = 1$:

$$E\left\{\mathfrak{L}_0\left(A_u\right)\right\} = \left[\frac{T^2\lambda_2}{(2\pi)^{3/2}} u + \frac{2T\lambda_2^{1/2}}{2\pi}\right] e^{\frac{-u^2}{2}} + \Psi\left(u\right).$$

Основной результат

Математическое ожидание Эйлеровой характеристики при $N = 2, \ \sigma^2 = 1, \ \lambda_2 = 200, \ \lambda_2 = 1000$:



$$\left| P\left\{ \sup_{x \in M} f\left(x\right) \geq u \right\} - E\left\{ \mathfrak{L}_0\left(A_u\left(f, M\right)\right) \right\} \right| \leq error(u).$$

$$\sup_{x \in M} f\left(x\right) \geq u \Leftrightarrow A_u\left(f, M\right) \neq \varnothing \Leftrightarrow \sharp\left\{\text{связные компоненты } A\right\} \geq 1$$

$$\sup_{x \in M} f\left(x\right) \geq u\text{“} \Leftrightarrow'' \mathfrak{L}_0\left(A_u\left(f, M\right)\right) = 1$$

$$P\left\{ \sup_{x \in M} f\left(x\right) \geq u \right\} \approx E\left\{ \mathfrak{L}_0\left(A_u\left(f, M\right)\right) \right\}$$

Для гауссова изотропного поля с постоянной единичной дисперсией $error\left(u\right) = e^{-u^2\left(1+\sigma_c^2\right)/2}, \quad \sigma_c^2 = \left.\frac{\partial^4 R(x)}{\partial x_1^2}\right|_{x=0} - 1.$

Другие результаты

Известны способы вычисления матожидания Эйлеровой и аналогичных характеристик в следующих условиях:

- f — достаточно гладкое, стационарное, A не касается границ, граница ∂M — C^2 многообразие;
- f — достаточно гладкое, стационарное, $N = 2,\ 3$, граница ∂M — C^2 многообразие за исключением конечного набора плоских граней и вершин, где они встречаются;
- f — достаточно гладкое, стационарное, изотропное, граница ∂M — C^2 многообразие за исключением конечного набора плоских граней и вершин, где они встречаются;
- f — гауссово или окологауссово, с постоянной дисперсией, M — расслоённое многообразие Уитни.

Достаточная гладкость: f — вещественнозначная, дважды непрерывно дифференцируемая в открытой окрестности M, не имеет критических точек на ∂M, сужения $f|_M$ и $f|_{\partial M}$ не имеют вырожденных критических точек.
Критическая точка x^* $\left(\frac{\partial f}{\partial x_i}(x^*) = 0,\ \ i = 1, \ldots, N\right)$ вырождена, если якобиан f в ней равен нулю.