My first scientific paper

Week 5

Highlight the principles

Vadim Strijov

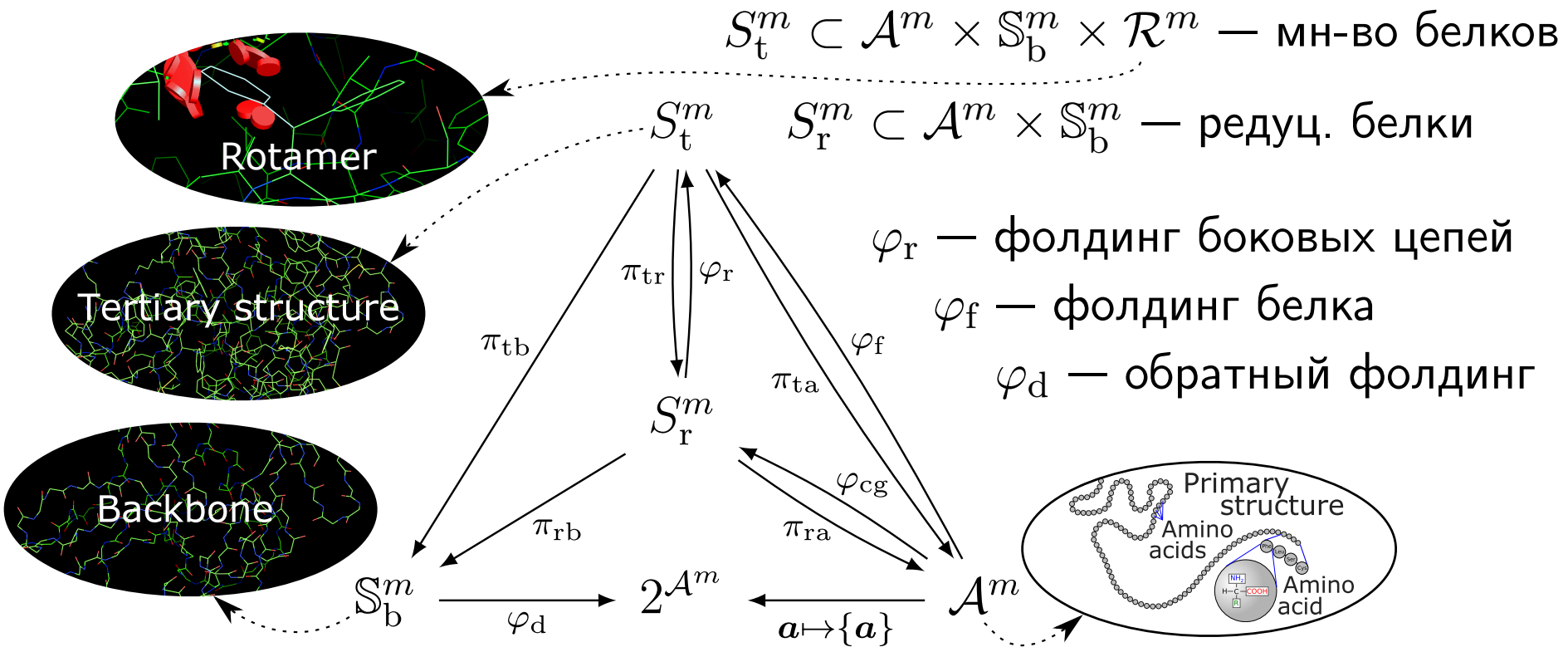
Moscow Institute of Physics and Technology

2021

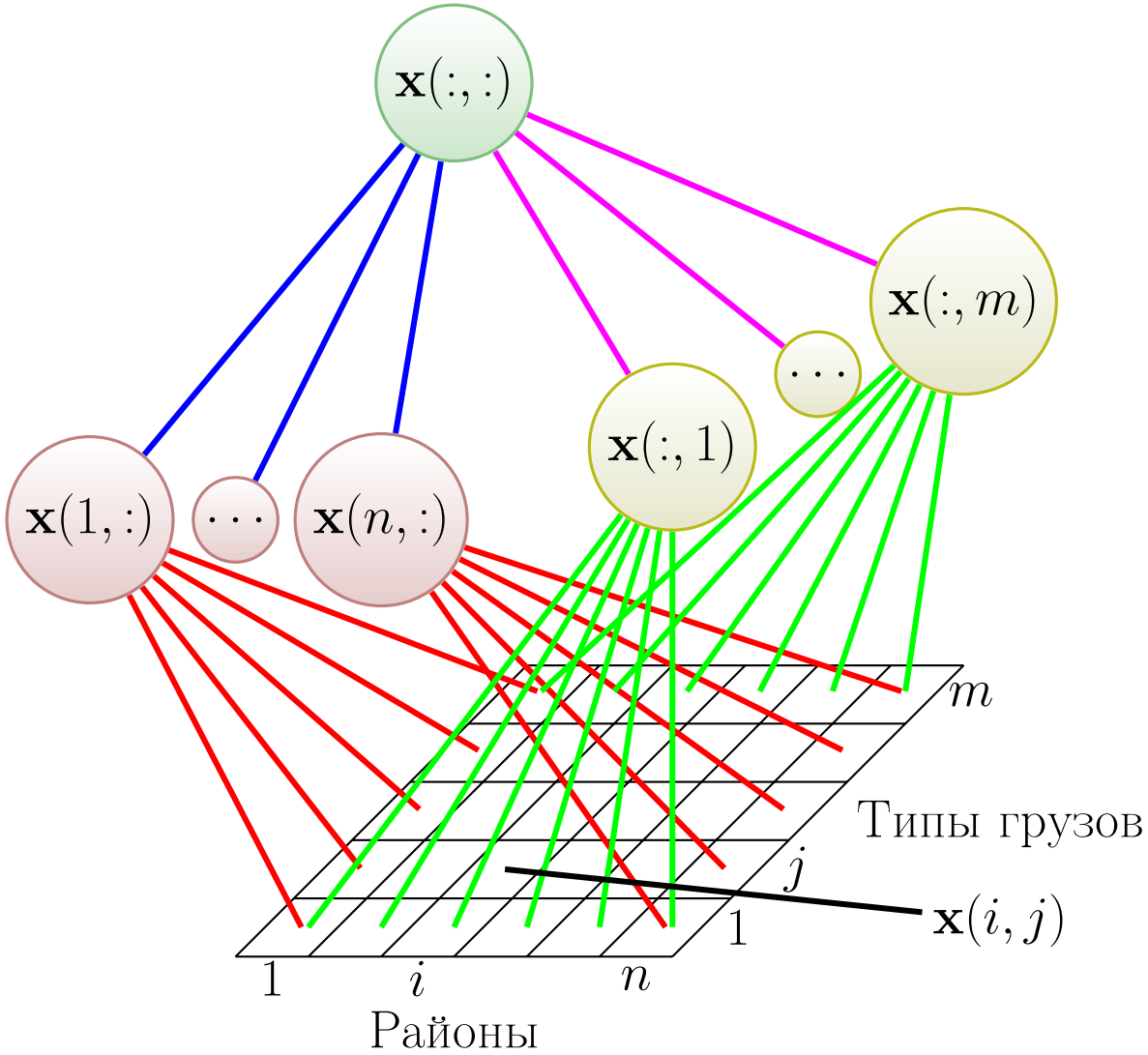
Задачи структурной биологии для белков длины m

$$\mathcal{A} = \{\mathsf{Ala}, \mathsf{Arg}, \mathsf{Asn}, \mathsf{Asp}, \mathsf{Cys}, \mathsf{Glu}, \mathsf{Gln}, \mathsf{Gly}, \mathsf{His}, \ldots, \mathsf{Trp}, \mathsf{Tyr}, \mathsf{Val}\}$$

$$\mathsf{Cys} = [\underbrace{\mathsf{N}, \mathsf{C}_\alpha, \mathsf{C}, \mathsf{H}, \mathsf{O}}_{\substack{\text{скелетная часть} \\ \mathbb{S}_\mathrm{b}=\mathbb{R}^{3\times 3}}}, \underbrace{\mathsf{H}_\alpha, \mathsf{C}_\beta, \mathsf{H}_{\beta_1}, \mathsf{H}_{\beta_2}, \mathsf{S}_\gamma, \mathsf{H}_\gamma}_{\substack{\text{боковая цепь} \\ \mathcal{R}}}]$$



Условие согласованности прогнозов



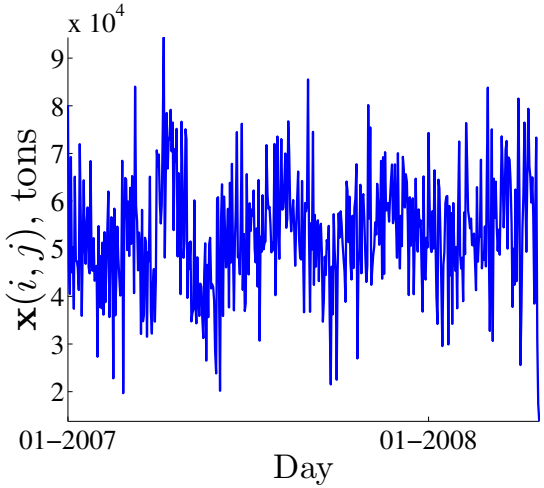
$$x_t(:,:) = \sum_{i=1}^{n} x_t(i,:);$$

$$x_t(:,:) = \sum_{j=1}^{m} x_t(:,j);$$

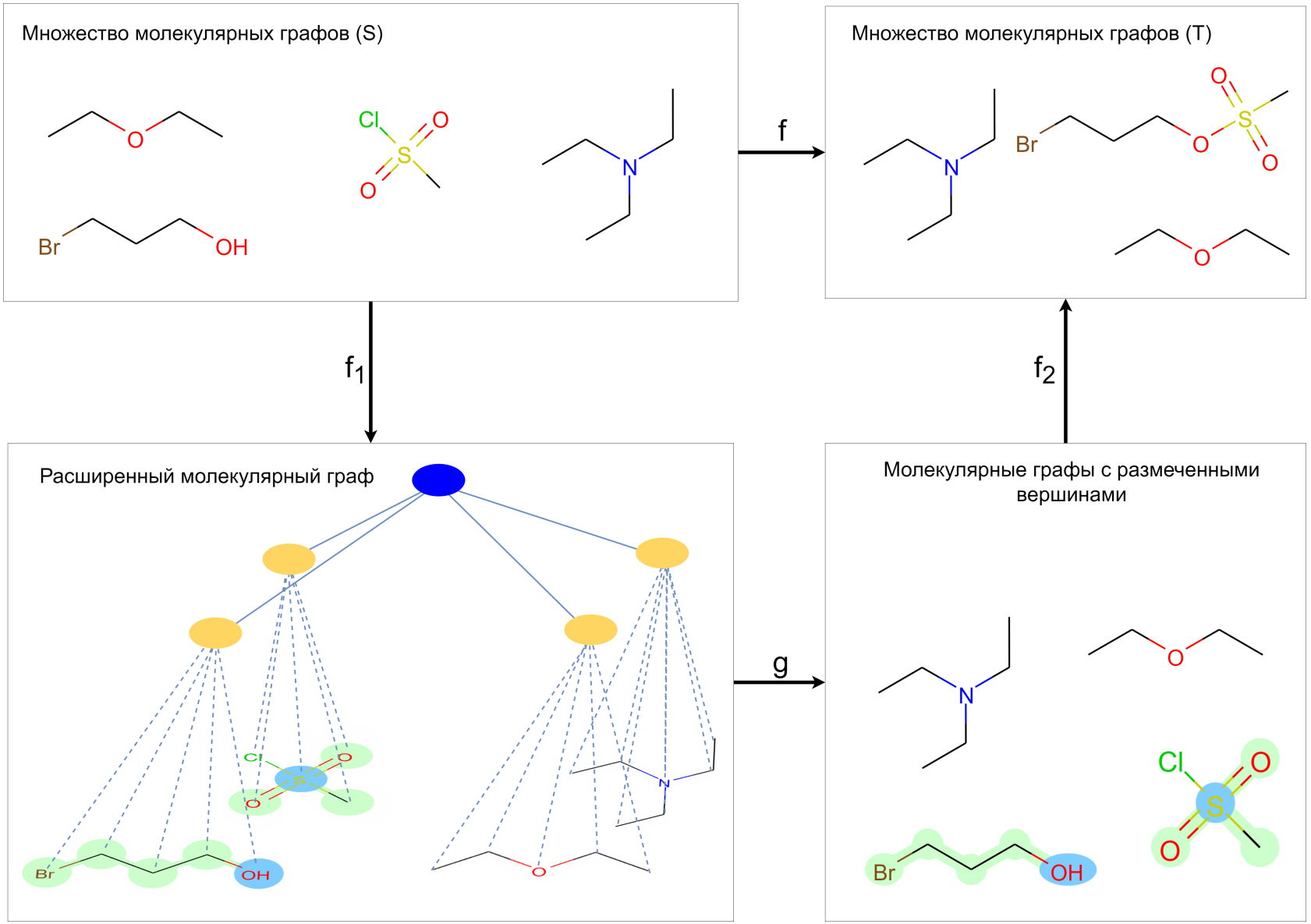
$$x_t(i,:) = \sum_{j=1}^{m} x_t(i,j), \quad i = 1,\dots n;$$

$$x_t(:,j) = \sum_{i=1}^{n} x_t(i,j), \quad j = 1,\dots m;$$

$$t = 1,\dots,T.$$



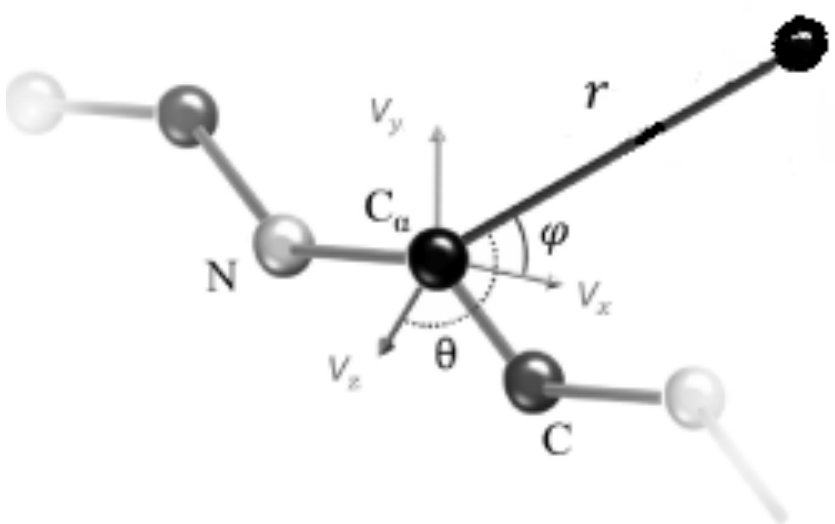
Структура решения



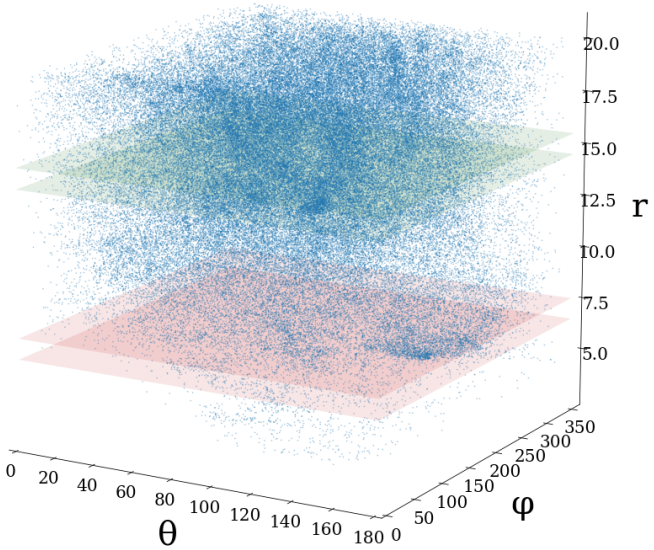
Описание молекулярной химической связи

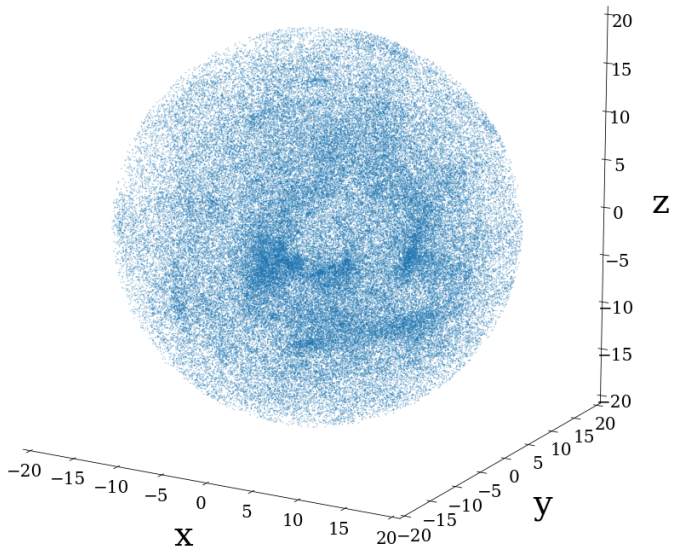
В данной работе исследуются взаимные пространственные ориентации различных пар молекул, образующих между собой химическую связь. Эта связь характеризуется тремя параметрами:

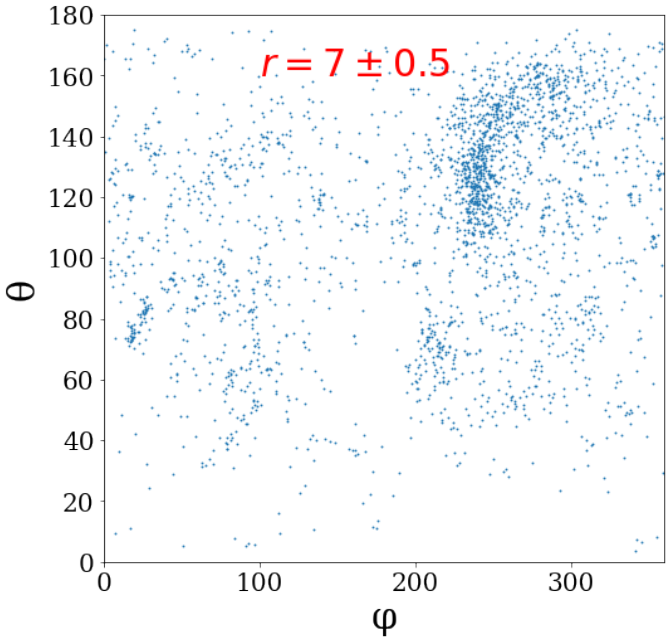
- r — расстояние между молекулами, $r \in [3\text{Å}, 20\text{Å}]$;
- (θ, φ) — пара сферических углов, определяющих положение лиганда в системе координат аминокислоты, $theta \in [0, \pi]$, $\varphi \in [0, 2\pi]$.

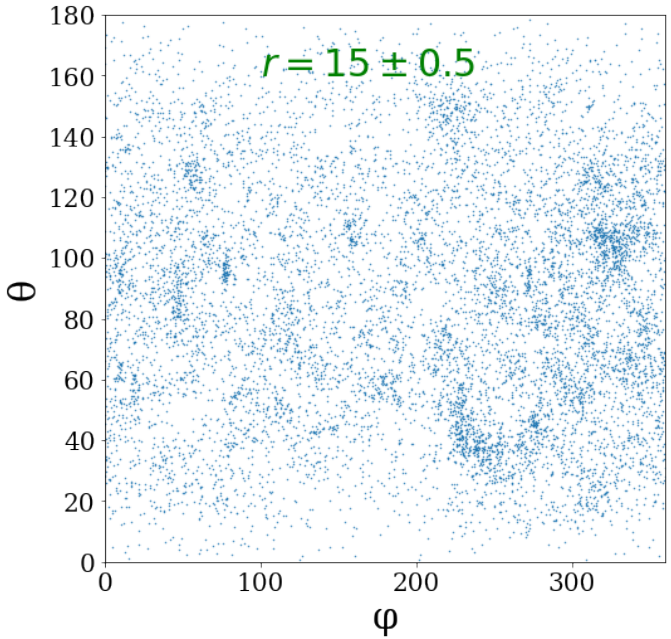


Представление выборки для пары ALA-C_{ar}



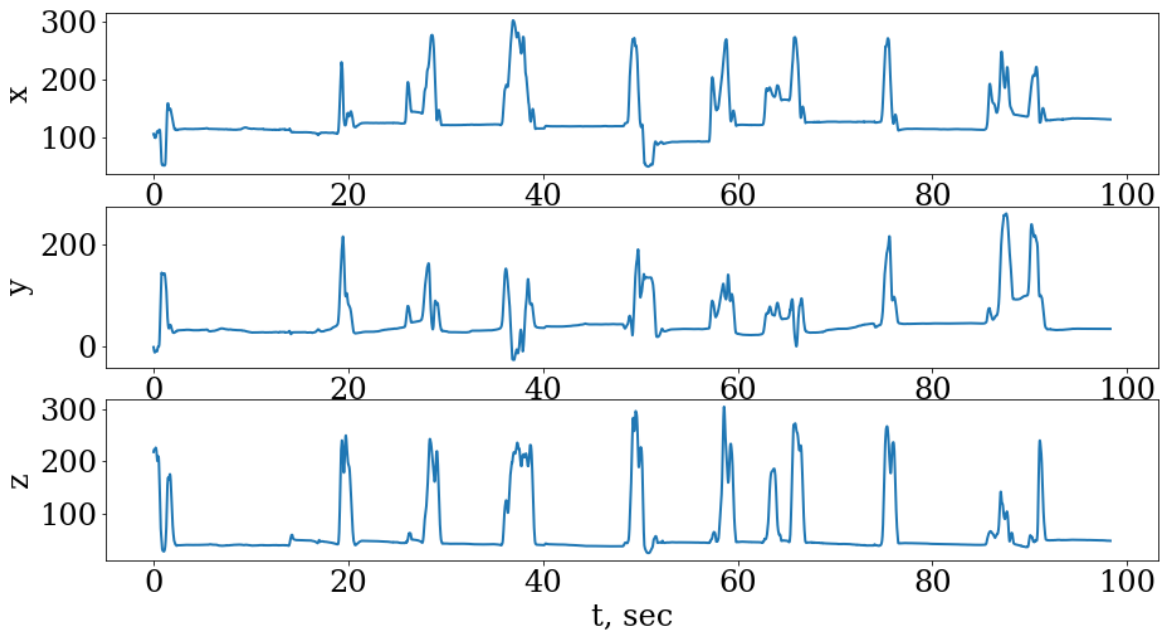




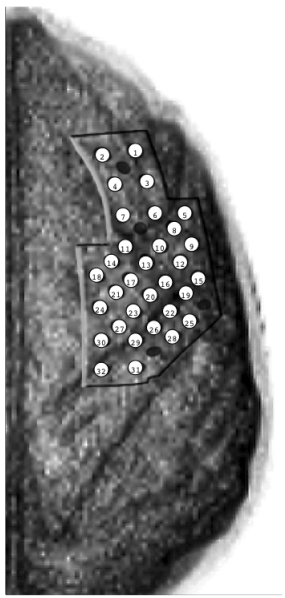


Декодируемые сигналы электрокортикограммы

- Сигналы $\mathbf{s}(t) \in \mathbb{R}^{N_{ch}}$. N_{ch} – число электродов
- Координаты электродов $\mathbf{Z} = \{(\mathbf{z}_j \in \mathbb{R}^2, j \in \{1 \ldots, N_{ch}\}\}$
- Положение кисти в пространстве $\mathbf{y}(t) \in \mathbb{R}^3$



Координата руки



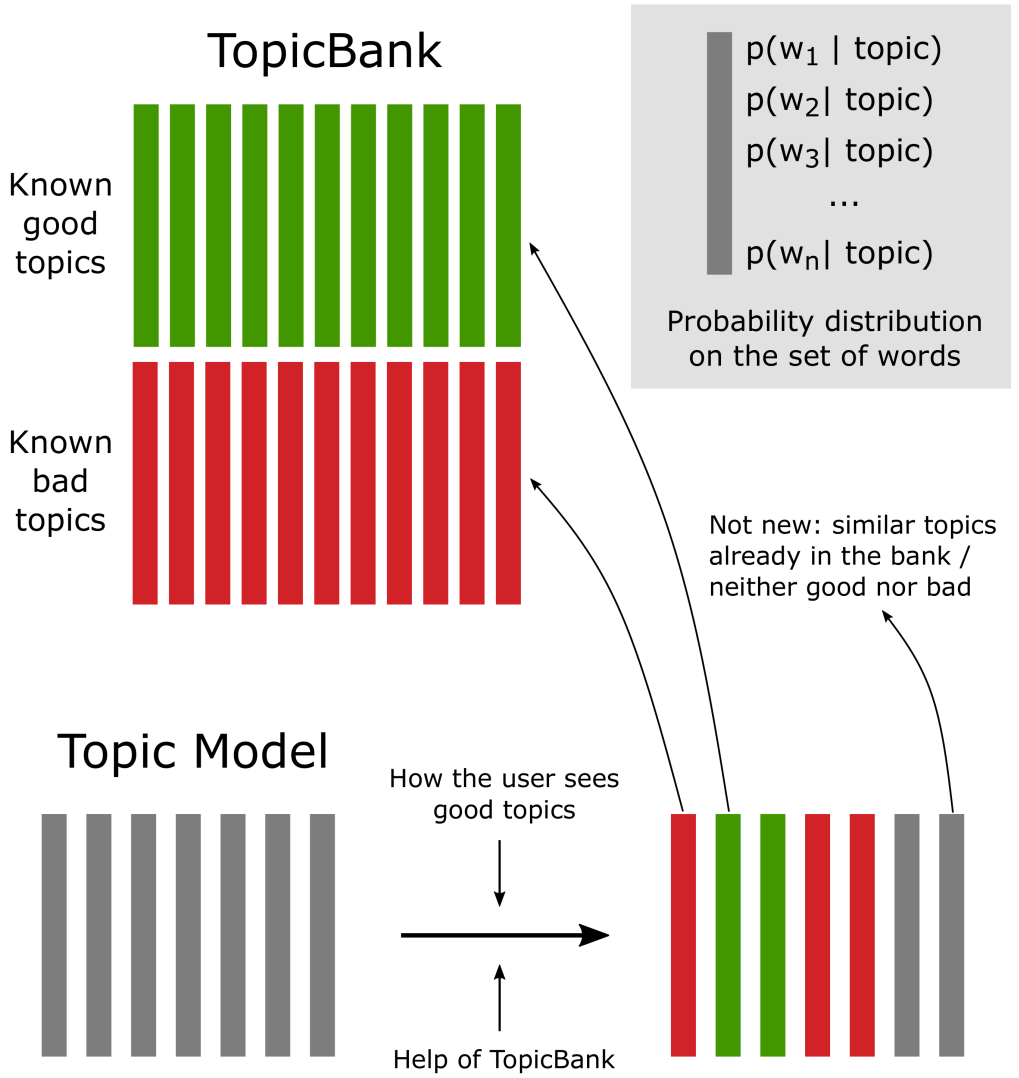
Пространственное расположение электродов

Chao ZC, Nagasaka Y, Fujii N (2010). "Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkeys."Frontiers in Neuroengineering 3:3.

Банк тем: сохранение интерпретируемых тем

Банк тем — модель полного набора тем: таких тем, которые

1) интерпретируемы,
2) существенно различны,
3) обеспечивают высокое правдоподобие модели $p(\Phi, \Theta \mid D)$.

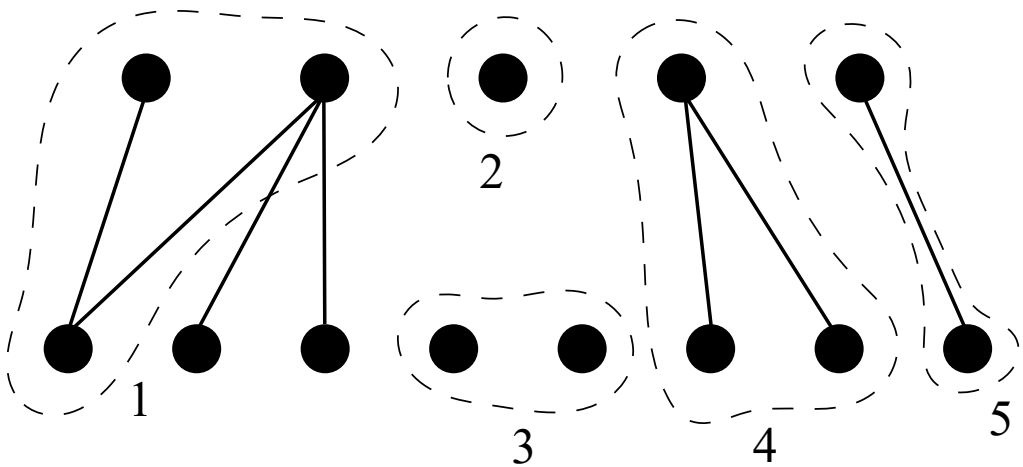


Построение банка тем

Аналогично построению двухуровневой иерархической тематической модели:

$$\underbrace{p(w \mid t)}_{\varphi_{wt}^{parent}} = \sum_{s \in S} \underbrace{p(w \mid s)}_{\varphi_{ws}^{child}} \underbrace{p(s \mid t)}_{\psi_{st}} \quad \text{Hierarchy}$$

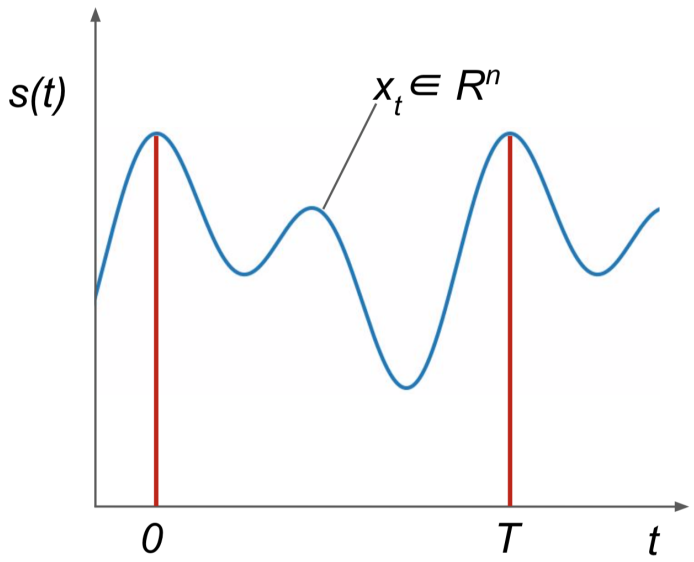
$$\underbrace{p(w \mid t)}_{\varphi_{wt}^{\mathbf{bank}}} = \sum_{s \in S} \underbrace{p(w \mid s)}_{\varphi_{ws}^{\mathbf{new}}} \underbrace{p(s \mid t)}_{\psi_{st}} \quad \text{TopicBank}$$

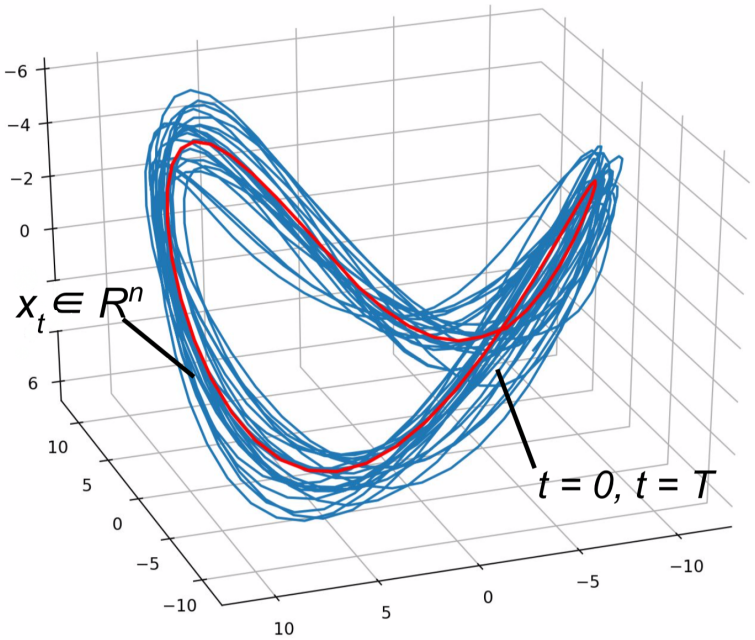


№	Hierarchy	TopicBank
1	ok	no
2	ok	ok
3	no	ok
4	ok	maybe
5	ok	maybe

Фазовая траектория

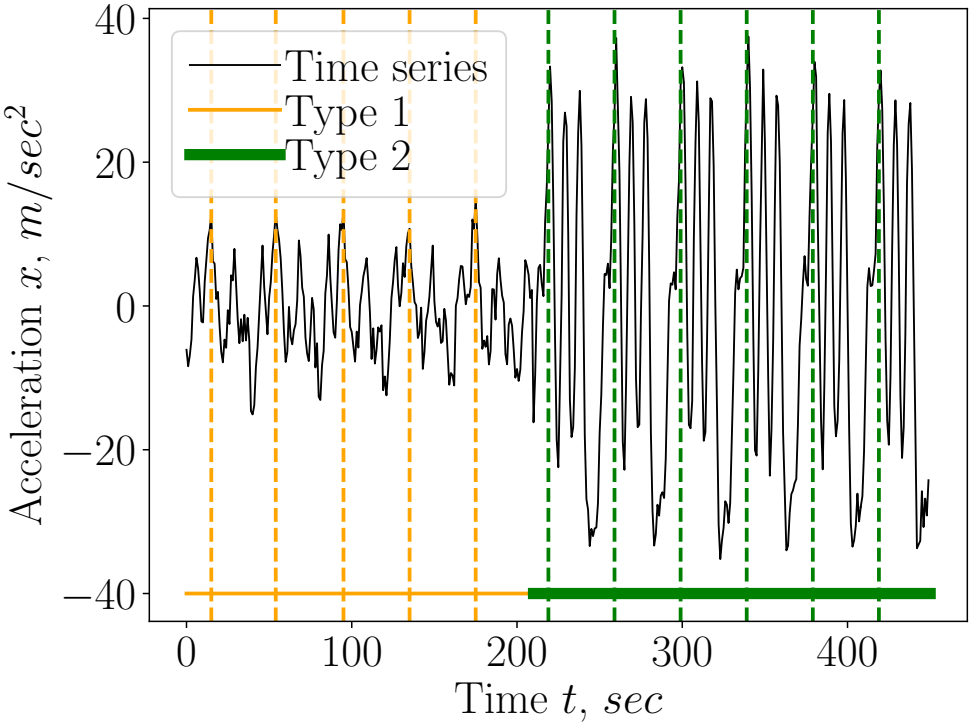
На рисунке представлен временной ряд и проекция его фазовой траектории в трехмерное пространство. $\mathbf{x}_t = \mathbf{x}(t)$ – точка на фазовой траектории в момент времени t.



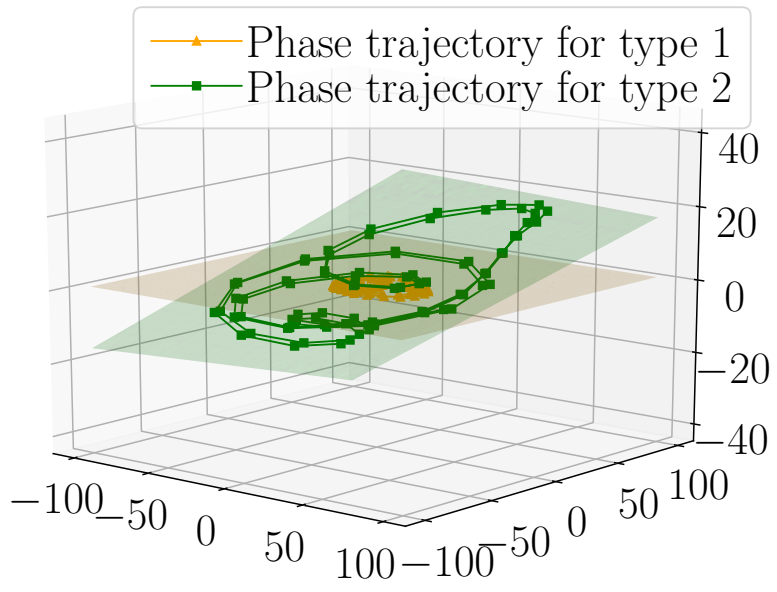


Постановка задачи кластеризации точек

Сегмент — последовательность точек временного ряда, которая относится к одному характерному физическому действию человека: шаг, прыжок.
Цепь — последовательность сегментов, которые образуют квазипериодическую последовательность точек.



(a)



(b)

a) временной ряда разбитый на сегменты; b) проекции на плоскость фазовых траекторий временного ряда, которые относятся к Type 1 и Type 2.

Пространство параметров

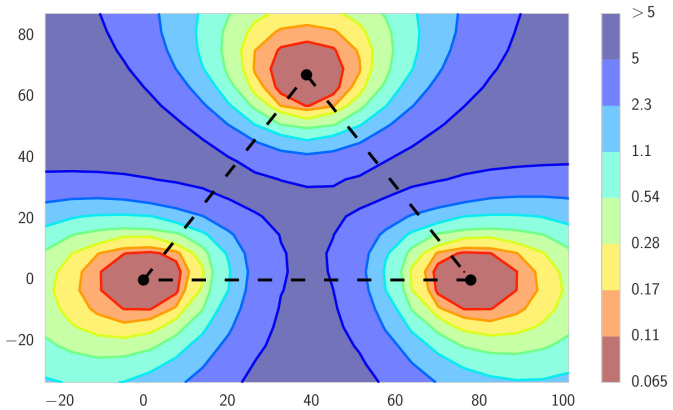


Рис.: изолированные оптимумы

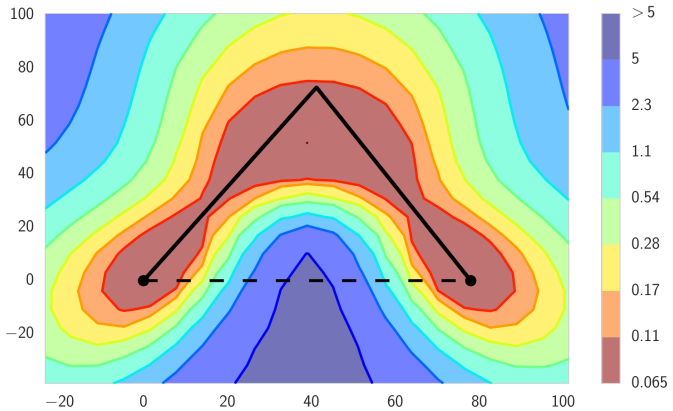
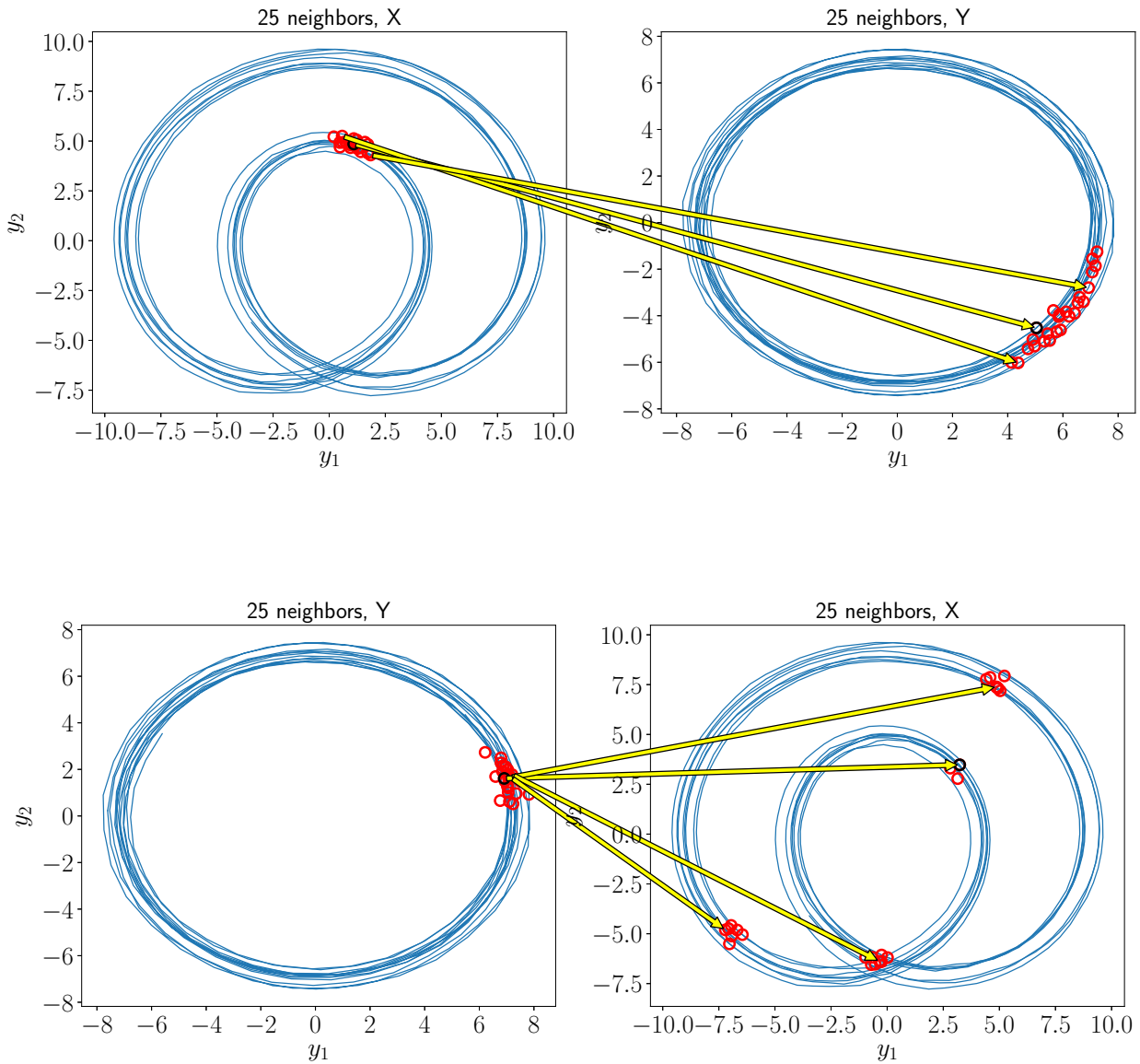


Рис.: не изолированные оптимумы

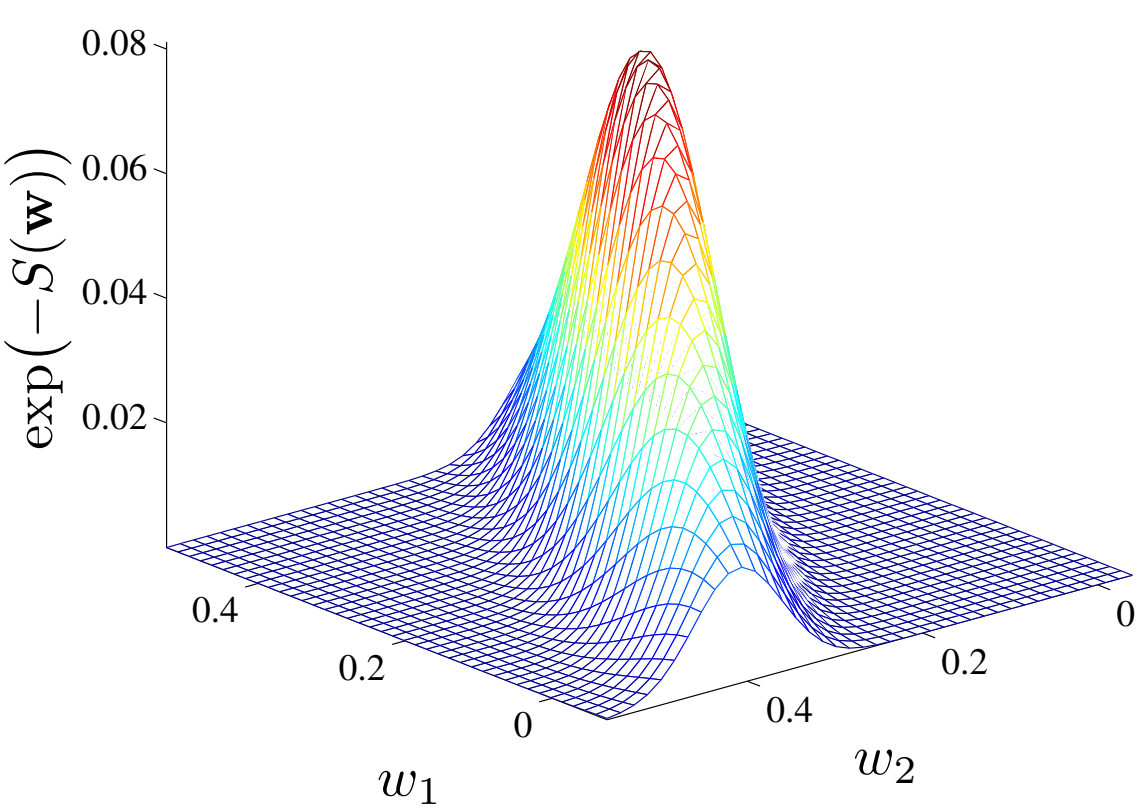
[1] https://arxiv.org/pdf/1802.10026.pdf

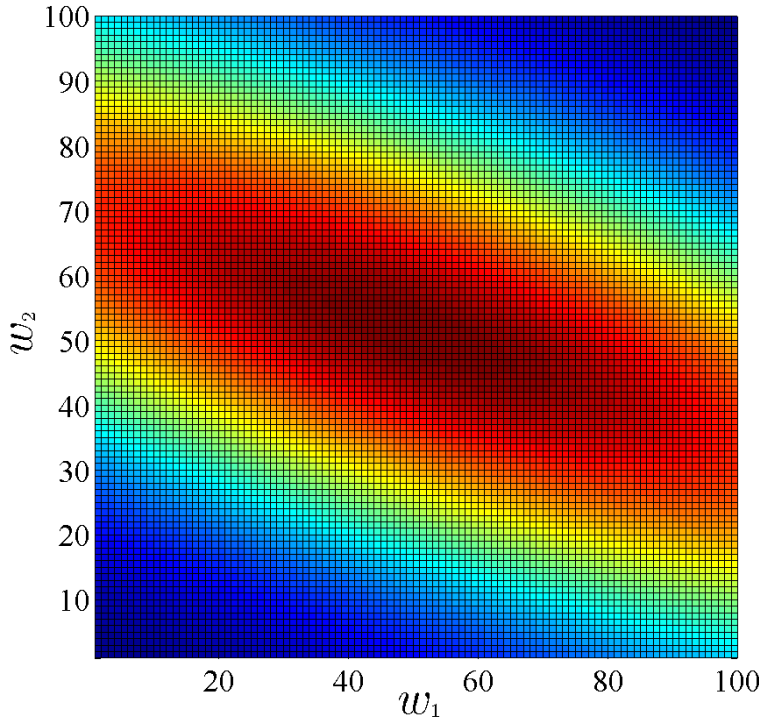
Ближайшие соседи на фазовых траекториях



Empirical distribution of model parameters

The value of error function $S(\mathbf{w}|\mathfrak{D}, f)$ depends on parameters.





x-axis and y-axis: parameters $\mathbf{w}$, z-axis: $\exp(-S(\mathbf{w}))$

Kuznetsov, Tokmakova, Strijov. 2016. Analytic methods of structure parameter // Informatica

Probabilistic model selection

Bayesian inference delivers the error function $S(\mathbf{w})$

Posterior Likelihood Prior

$$p(\mathbf{w}|\mathfrak{D},\mathbf{A},\mathbf{B},\mathbf{f}) = \frac{p(\mathfrak{D}|\mathbf{w},\mathbf{B},\mathbf{f})p(\mathbf{w}|\mathbf{A},\mathbf{f})}{p(\mathfrak{D}|\mathbf{A},\mathbf{B},\mathbf{f})}.$$

Evidence (to select a model)

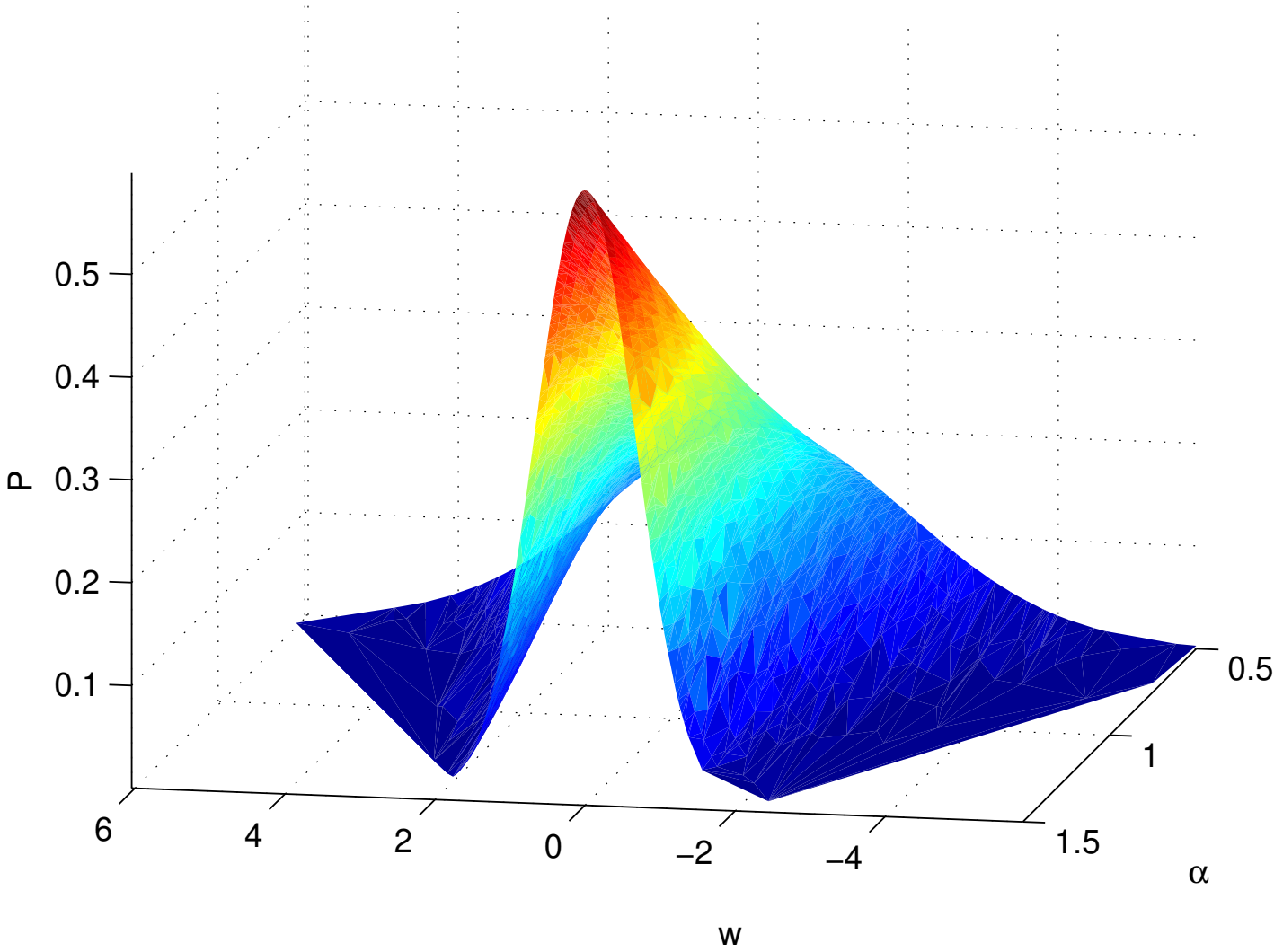
Write the error function given hyperparameters $\mathbf{A}, \mathbf{B}$

$$S(\mathbf{w}) = \underbrace{\frac{1}{2}(\mathbf{y}-\mathbf{f})^{\mathsf{T}}\mathbf{B}(\mathbf{y}-\mathbf{f})}_{\text{approximation error}} + \underbrace{\frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^{\mathsf{T}}\mathbf{A}(\mathbf{w}-\hat{\mathbf{w}})}_{\text{regularisation error}},$$

$$S = E_D + E_{\mathbf{w}} = \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{s}, \qquad \text{metaparameters } \lambda = \frac{1}{2}.$$

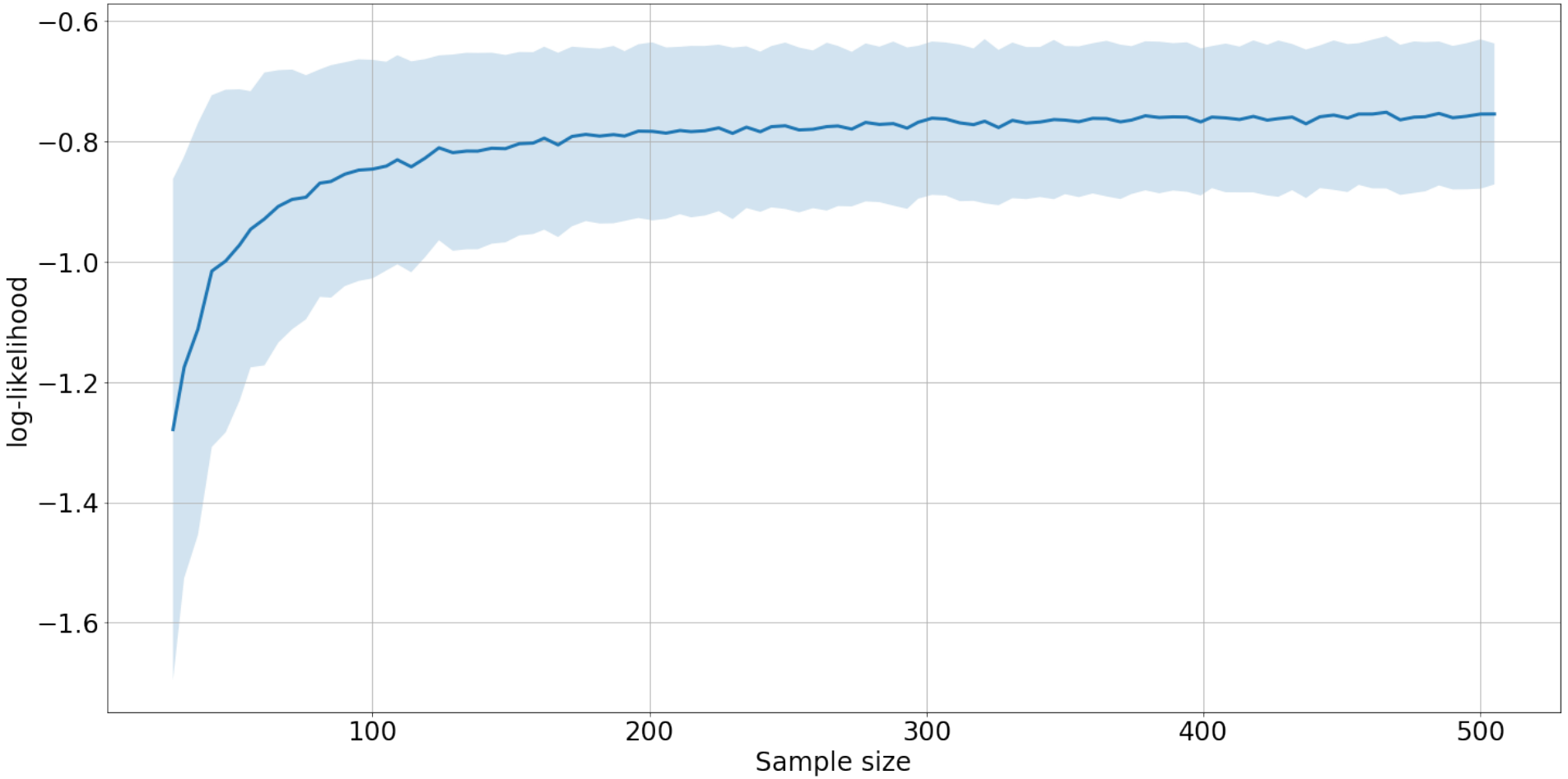
Evidence of the model

depends on both, error E_D (likelihood) and regularisation $E_{\mathbf{w}}$ (prior).

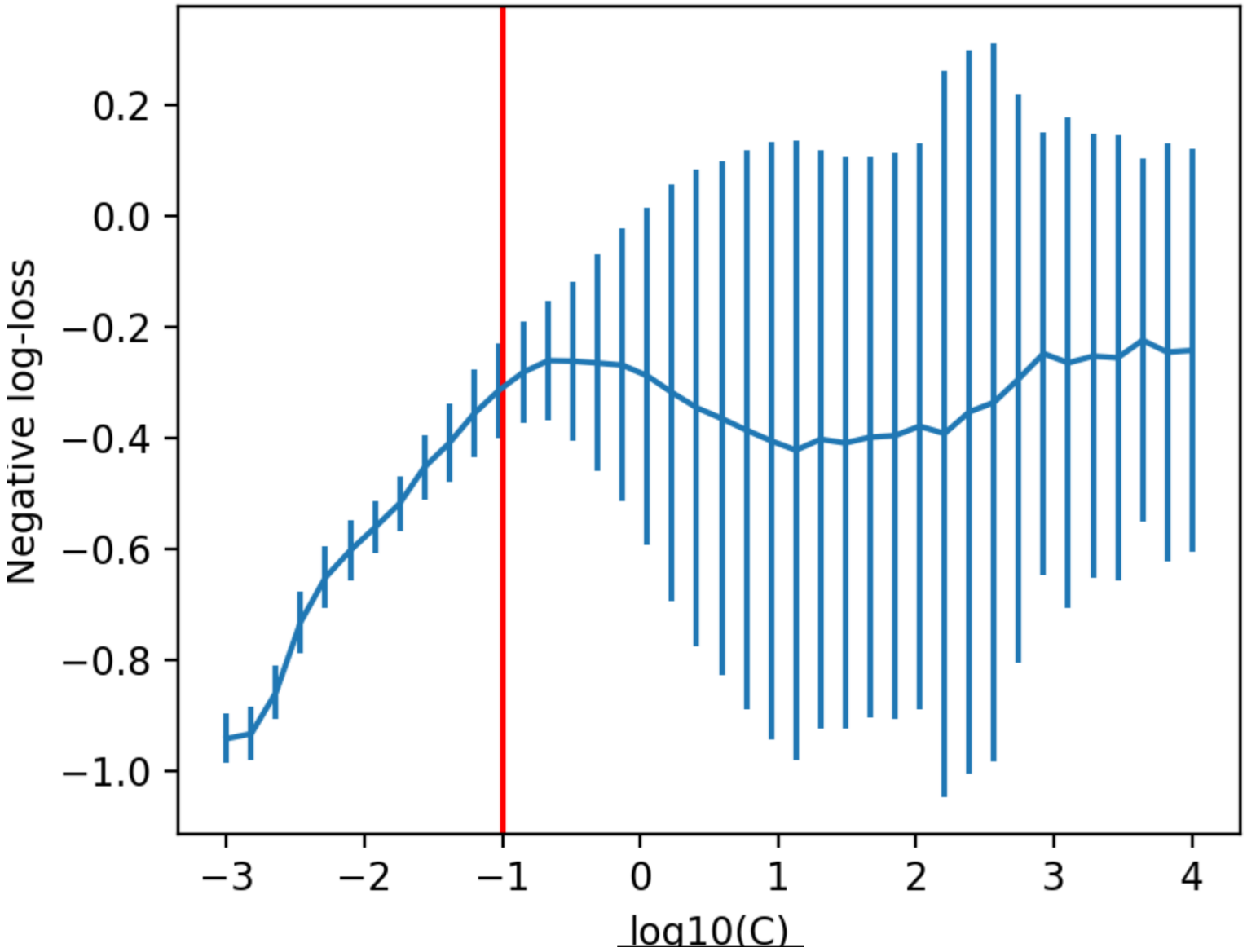


Parameters $\mathbf{w}$, variance α^{-2}, and $p(\mathbf{w}|\mathfrak{D}, \alpha)$ is the evidence.

– Error and its variance for a reinforced sample set



Variance of error increasing over model complexity



Complexity, number of model parameters

Гипотеза порождения данных для линейной модели

Пусть $\mathbb{E}(\mathbf{y}|X) = \mathbf{f}$ и многомерная случайная величина имеет нормальное распределение

$$p(\mathbf{y}) = (2\pi)^{-\frac{m}{2}} \det{}^{-\frac{1}{2}} \left(B^{-1}\right) \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{f})^T B(\mathbf{y}-\mathbf{f})\right).$$

Рассмотрим три варианта. Элементы вектора $\mathbf{y}$ имеют

1) одинаковую дисперсию и независимы, $\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_l) = 0, i \neq l$,
$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \beta^{-1} I),$$
2) имеют различную дисперсию и независимы,
$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \mathrm{diag}(\beta_1, \ldots, \beta_m)^{-1} I)$$
3) описываются ковариационной матрицей общего вида,
$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, B^{-1});$$
эта матрица симметрична и положительно определена.

Функция правдоподобия данных

Функция вероятности появления зависимой переменной имеет вид

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, B, f) \stackrel{\text{def}}{=} p(D|\mathbf{w}, \beta, f) = \frac{\exp(-E_D)}{Z_D(B)}.$$

Функция ошибки, соответствующая математическому ожиданию регрессионной модели при данной гипотезе, определена как

$$E_D = \frac{1}{2}(\mathbf{y} - \mathbf{f})^T B(\mathbf{y} - \mathbf{f}).$$

Коэффициент Z_D определен выражением, нормирующим функцию плотности нормального распределения

$$Z_D(B) = (2\pi)^{\frac{m}{2}} \det{}^{\frac{1}{2}} \left(B^{-1}\right).$$

Функция правдоподобия данных при $B = \beta I$

Для гомоскедаксичного случая функция ошибки равна

$$E_D = \frac{1}{2}\beta \sum_{i \in \mathcal{I}} (y_i - f(\mathbf{w}, \mathbf{x}_i))^2 ,$$

а нормирующий множитель

$$Z_D(\beta) = \left(\frac{2\pi}{\beta}\right)^{\frac{m}{2}} .$$

Априорное (sic!) распределение параметров модели

Из принятой гипотезы порождения данных следует нормальность распределения параметров, $\mathbf{w} \sim \mathcal{N}(\mathbf{w_0}, A^{-1})$:

$$p(\mathbf{w}|A, f) = \frac{\exp(-E_\mathbf{w})}{Z_\mathbf{w}(A)}.$$

Функция-штраф за большое значение параметров модели для принятого распределения определена как

$$E_\mathbf{w} = \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T A(\mathbf{w} - \mathbf{w}_0).$$

Нормирующая константа $Z_\mathbf{w}$ равна

$$Z_\mathbf{w}(A) = (2\pi)^{\frac{n}{2}} \det{}^{\frac{1}{2}}(A^{-1}).$$

При равенстве дисперсий элементов вектора параметров

$$Z_\mathbf{w}(\alpha) = \left(\frac{2\pi}{\alpha}\right)^{\frac{m}{2}} \quad \text{и} \quad E_\mathbf{w} = \frac{1}{2}\alpha\|\mathbf{w}\|^2.$$

Байесовский вывод, первый уровень

Апостериорное распределение параметров модели для заданных матриц A, B имеет вид

$$p(\mathbf{w}|D, A, B, f) = \frac{p(D|\mathbf{w}, B, f)p(\mathbf{w}|A, f)}{p(D|A, B, f)}.$$

Элементы этого выражения и соответствующие им параметры:

- $p(\mathbf{w}|D, A, B, f)$ — апостериорное распределение параметров,
- $\mathbf{w}_{\text{MP}} = \arg\max p(\mathbf{w}|D, A, B, f)$ — наиболее вероятные параметры,
- $p(D|\mathbf{w}, B, f)$ — функция правдоподобия данных,
- $\mathbf{w}_{\text{ML}} = \arg\max p(D|\mathbf{w}, B, f)$ — наиболее правдоподобные параметры,
- $p(\mathbf{w}|A, f)$ — априорное распределение параметров,
- $p(D|A, B, f)$ — функция правдоподобия модели.

Апостериорное распределение параметров, частный случай

Апостериорное распределение параметров модели для заданных матриц A, B

$$p(\mathbf{w}|D, A, B, f) = \frac{p(D|\mathbf{w}, B, f)p(\mathbf{w}|A, f)}{p(D|A, B, f)}.$$

Записывая функцию ошибки $S = E_{\mathbf{w}} + E_D$ в виде

$$S(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T A(\mathbf{w} - \mathbf{w}_0) + \frac{1}{2}(\mathbf{y} - \mathbf{f})^T B(\mathbf{y} - \mathbf{f}),$$

получаем вместо вышестоящего выражение

$$p(\mathbf{w}|D, A, B, f) \propto \frac{\exp(-S(\mathbf{w}))}{Z_S},$$

где Z_S — нормирующий множитель.

Апостериорное распределение параметров, частный случай

При рассмотрении частных случаев ковариационных матриц $B = \beta I_m$ и $A = \alpha I_n$ и при $\mathbf{w}_0 = \mathbf{0}$ апостериорное распределение параметров принимает вид

$$p(\mathbf{w}|D, \alpha, \beta, f) = \frac{p(D|\mathbf{w}, \beta, f)p(\mathbf{w}|\alpha, f)}{p(D|\alpha, \beta, f)}.$$

а функция ошибки —

$$S(\mathbf{w}) = \frac{1}{2}\alpha\|\mathbf{w}\|^2 + \frac{1}{2}\beta\|\mathbf{y} - \mathbf{f}\|^2.$$

Параметры α и β в последнем выражении играют роль регуляризирующих множителей.

Функция ошибки включает две матрицы ковариации

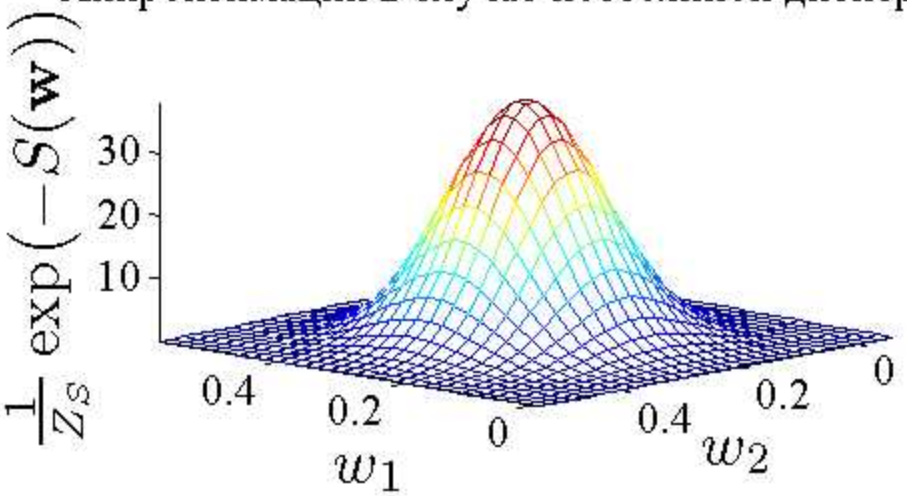
Согласно первому уровню Байесовского вывода

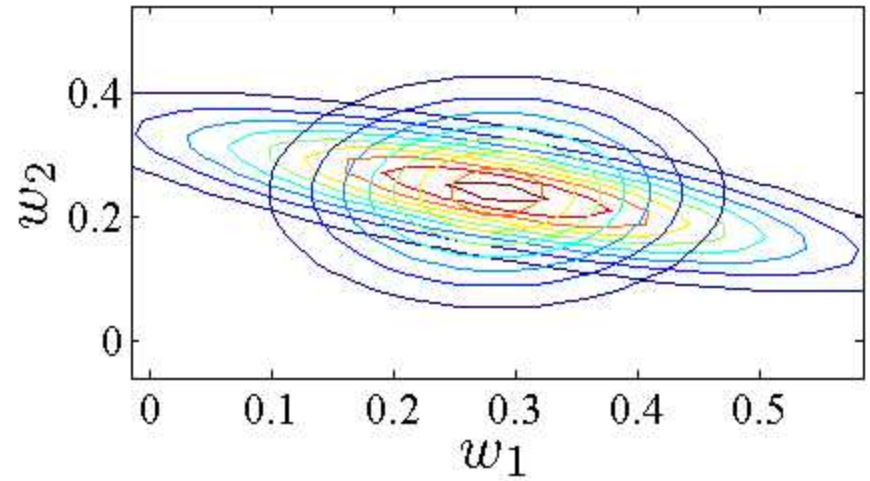
$$S(\mathbf{w}|D, f) = \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MP}})^{\mathsf{T}} A(\mathbf{w} - \mathbf{w}_{\text{MP}}) + \frac{1}{2}(\mathbf{f} - \mathbf{y})^{\mathsf{T}} B(\mathbf{f} - \mathbf{y}).$$

Имеется девять возможных вариантов гипотезы порождения данных.

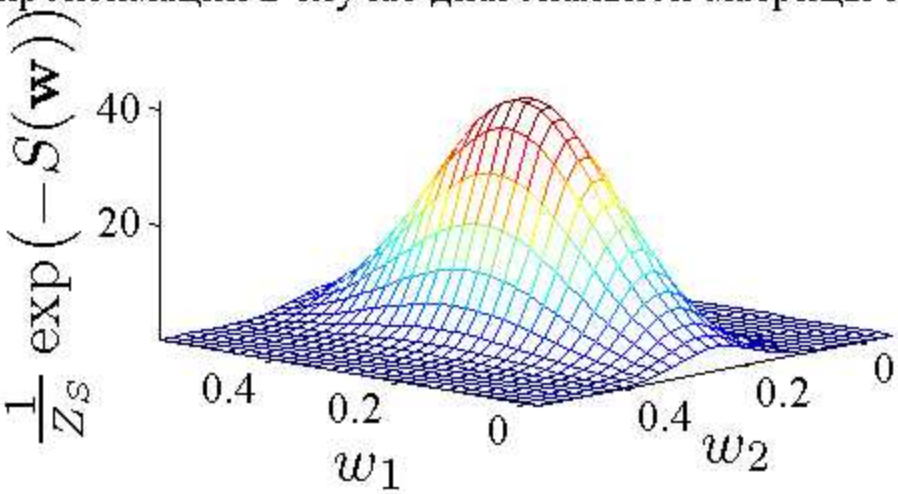
Обратная ковариационная матрица параметров	Обратная ковариационная матрица зависимой переменной
$A = \alpha I_n$	$B = \beta I_m$
$A = \text{diag}(\alpha_1, \ldots, \alpha_n)$	$B = \text{diag}(\beta_1, \ldots, \beta_m)$
A	B

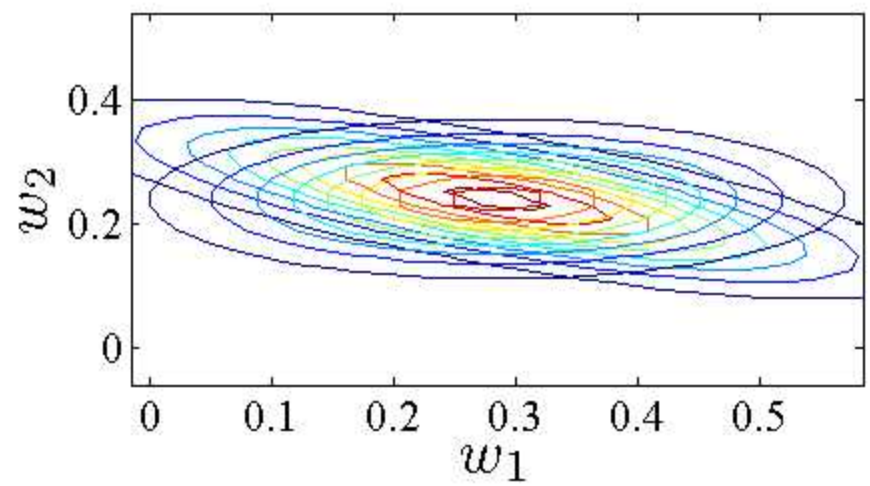
Аппроксимации в случае постоянной дисперсии





ппроксимации в случае диагональной матрицы ковариаций





ппроксимации в случае матрицы ковариаций общего вида