My first scientific paper

Week 5

Highlight the principles

Vadim Strijov

Moscow Institute of Physics and Technology

2021

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

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

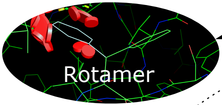
$S^m_{\mathrm{t}} \subset \mathcal{A}^m \times \mathbb{S}^m_{\mathrm{b}} \times \mathcal{R}^m$ — мн-во белков

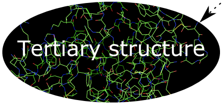
$S^m_{\mathrm{r}} \subset \mathcal{A}^m \times \mathbb{S}^m_{\mathrm{b}}$ — редуц. белки

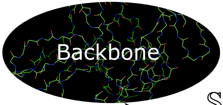
φ_{r} — фолдинг боковых цепей

φ_{f} — фолдинг белка

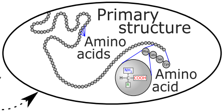
φ_{d} — обратный фолдинг



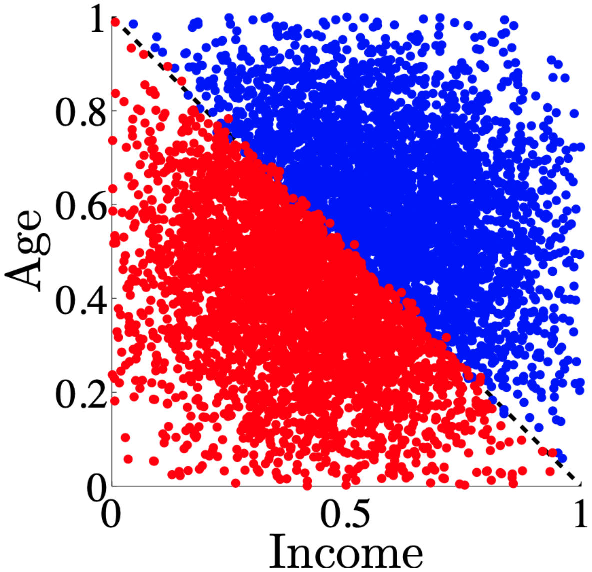




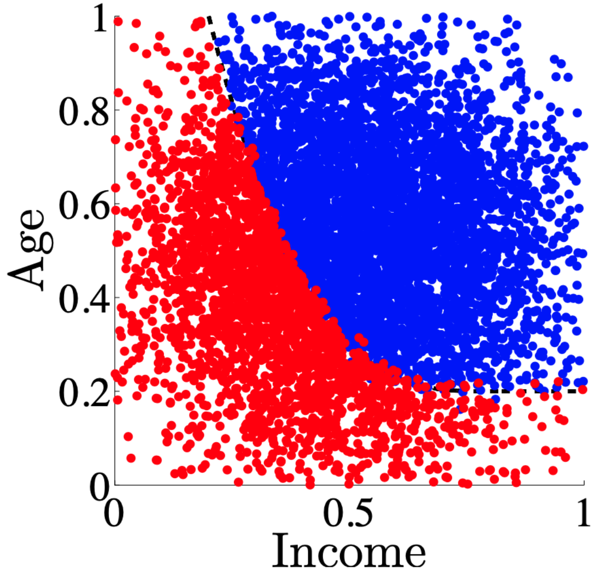
$$\begin{tikzcd}
 & & S^m_{\mathrm{t}} \arrow[ddll, "\pi_{\mathrm{tb}}"'] \arrow[d, shift right, "\pi_{\mathrm{tr}}"'] \arrow[ddrr, shift right, "\pi_{\mathrm{ta}}"'] & & \\
 & & S^m_{\mathrm{r}} \arrow[u, shift right, "\varphi_{\mathrm{r}}"'] \arrow[dll, "\pi_{\mathrm{rb}}"] \arrow[drr, shift right, "\pi_{\mathrm{ra}}"'] & & \\
\mathbb{S}^m_{\mathrm{b}} \arrow[rr, "\varphi_{\mathrm{d}}"'] & & 2^{\mathcal{A}^m} & & \mathcal{A}^m \arrow[ll, "\boldsymbol{a} \mapsto \{\boldsymbol{a}\}"] \arrow[uull, shift right, "\varphi_{\mathrm{f}}"'] \arrow[ull, shift right, "\varphi_{\mathrm{cg}}"']
\end{tikzcd}$$



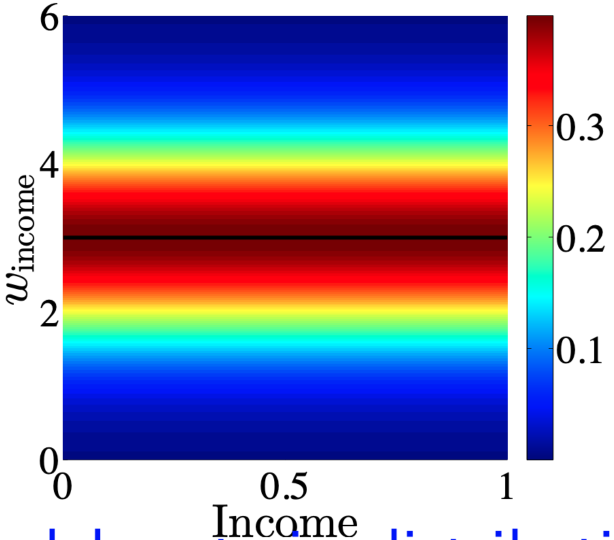
When one needs a multimodel?



Sample generation assumption

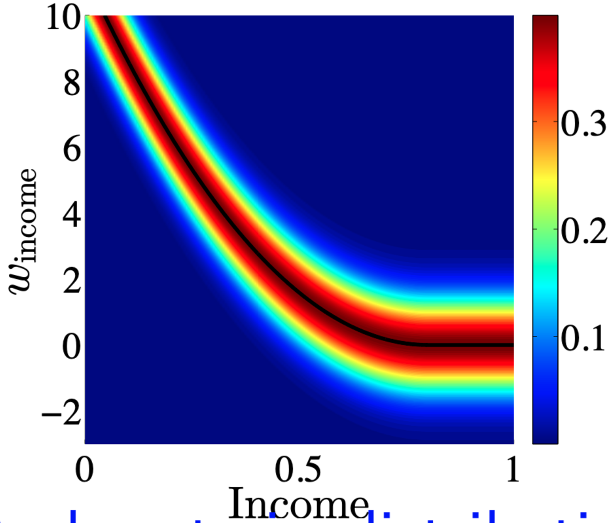


Real data



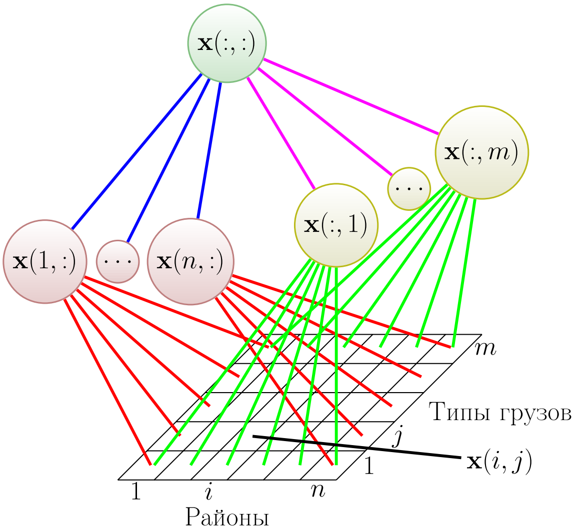
Model posterior distribution

$p(\mathbf{w}|\mathbf{X},\ \mathbf{y})$

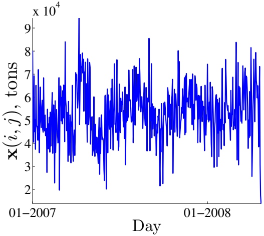


Real posterior distribution

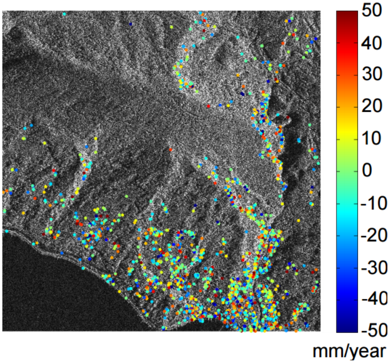
$p(\mathbf{w}|\mathbf{X},\ \mathbf{y})$



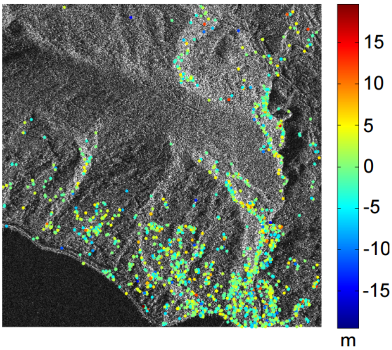
$$x_t(:,:) = \sum_{i=1}^{n} x_t(i,:);$$
$$x_t(:,:) = \sum_{j=1}^{m} x_t(:,j);$$
$$x_t(i,:) = \sum_{j=1}^{m} x_t(i,j), \quad i = 1,\dots n;$$
$$x_t(:,j) = \sum_{i=1}^{n} x_t(i,j), \quad j = 1,\dots m;$$
$$t = 1,\dots,T.$$



Анализ спутниковых снимков



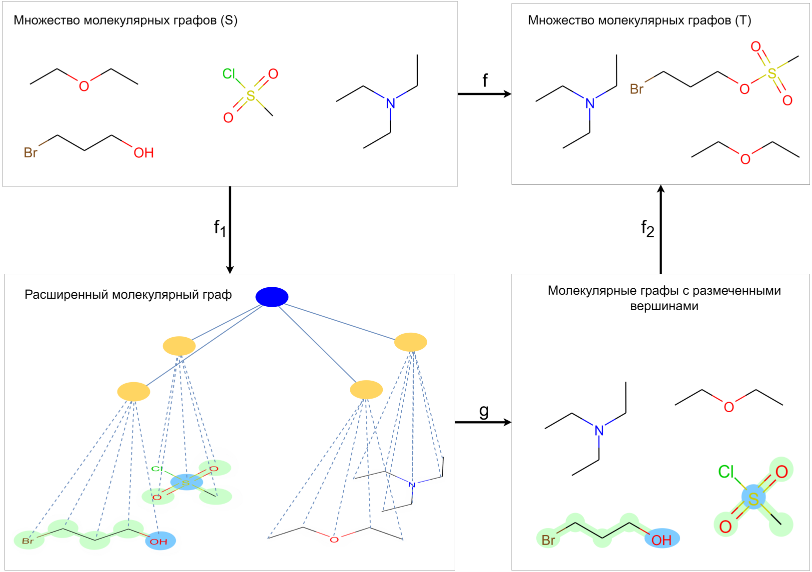
Скорость движения вдоль направления наблюдения



Погрешность в высоте

Рудаков К.В., Адуенко А.А., Рейер И.А., Василейский А.С., Карелов А.И., Стрижов В.В. Алгоритмы выделения и совмещения устойчивых отражателей на спутниковых снимках // Компьютерная оптика, 2015.

Структура решения



Задача регрессии в пространстве мультииндексных признаков

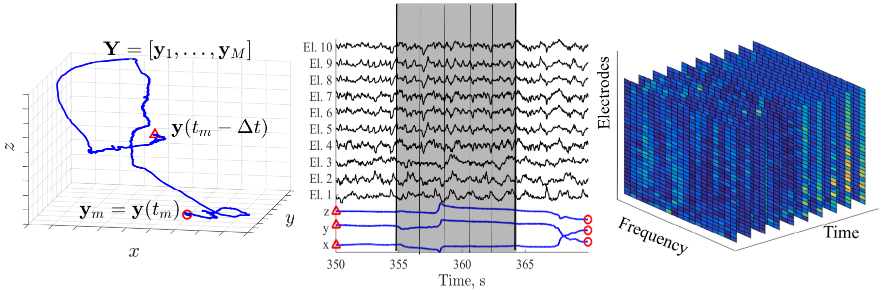
Решается задача восстановления регрессии

$$\hat{\mathbf{y}} = \sum_{i_1,\dots,i_D} \left[\underline{\mathbf{X}} * \underline{\mathbf{W}}\right]_{i_1\dots i_D}, \quad \underline{\mathbf{W}} = \arg\min_{\mathcal{W}} \sum_{m=1}^{M} \|\underline{\mathbf{X}}_m * \underline{\mathbf{W}} - \mathbf{y}_m\|_F^2.$$

где $\mathbf{y} \in \mathbb{R}^n$, $\underline{\mathbf{X}}$, $\underline{\mathbf{W}} \in \mathbb{R}^{n_1 \times \dots \times n_D}$ — многоиндексные матрицы, $*$ обозначает операцию поэлементного умножения.

В векторном виде задача имеет вид $\mathbf{y} = \mathbf{x}^\mathsf{T}\mathbf{w}$, где

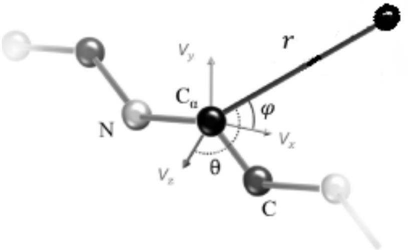
$$\mathbf{x} = \text{vec}\left(\underline{\mathbf{X}}\right) = x_{i_1(n_1-1)+\dots+i_{D-1}(n_{D-1}-1)+i_D} = [\underline{\mathbf{X}}]_{i_1\dots i_D}.$$



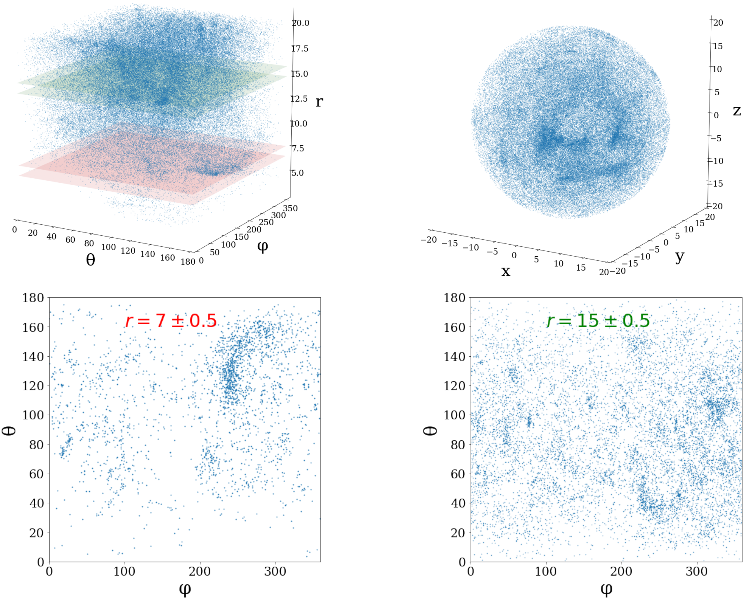
Описание молекулярной химической связи

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

- r — расстояние между молекулами, $r \in [3Å, 20Å]$;
- (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}

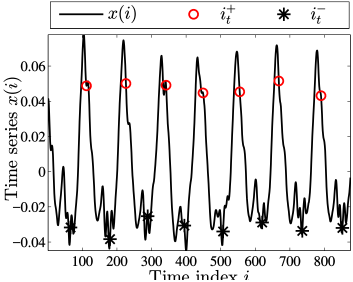


Сегментирование квазипериодического временного ряда

Квазипериодический временной ряд $\mathbf{s} = \{s(t_1), \dots, s(t_i), \dots, s(t_m)\}$ длины m определяется набором $\langle \mathbf{s}^*, a(i,s), f(i) \rangle$, так что

$$s(t_i) = a(i, s_{[f(i)]}),$$

где $\mathbf{s}^* = [s_1, \dots, s_T]^\mathsf{T}$ — базовый сегмент,
$a(i,s), i \in \{1, \dots, m\}$ — трансформация формы базового сегмента, $f(i) \mapsto \{1, \dots, T\}$ — масштабирование по времени.



Теорема (Мотренко)

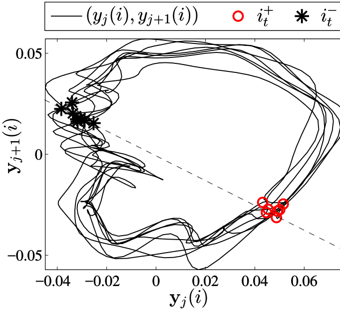
Для временного ряда $\mathbf{s}$ *вида* $s(t_i) = A_i \cos(2\pi w i + \phi)$ *с* $w \in (0, 1/2)$, $\phi \in [0, 2\pi)$, $m \cdot w \in \mathbb{N}$ *и* $A_i : \exists C \in \mathbb{R} |A_i| < C \, \forall i$ *главные компоненты* $\mathbf{y}_1$ *и* $\mathbf{y}_2$ *могут быть представлены в виде*

$$y_1(l) = B_1(l) \cos(2\pi w l + \phi_1),$$

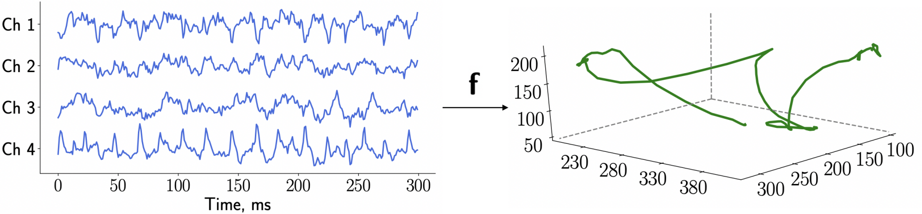
$$y_2(l) = B_2(l) \cos(2\pi w l + \phi_2),$$

$$\phi_1, \ \phi_2 \in [0, 2\pi), \ l = 1, \dots, m - N + 1$$

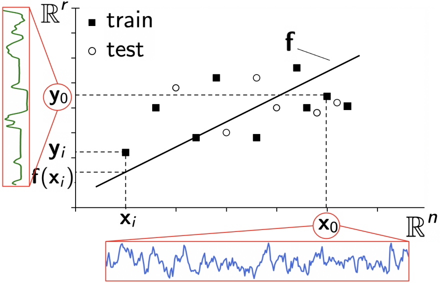
где разница между $|\phi_1 - \phi_2| \to \pi/2$.



Восстановление зависимости в исходном и целевом пространствах



Прогностическая модель декодирования



Согласование зависимостей в скрытом пространстве

$$\begin{tikzcd}
\mathbf{x} \in \mathbb{R}^n \arrow[rr, "\mathbf{f}"] \arrow[dr, bend right, "\mathbf{W}"'] & & \mathbf{y} \in \mathbb{R}^r \arrow[dl, bend left, "\mathbf{C}"] \\
& \mathbf{t}, \mathbf{u} \in \mathbb{R}^\ell \arrow[ul, bend right, "\mathbf{P}"'] \arrow[ur, bend left, "\mathbf{Q}"] &
\end{tikzcd}$$

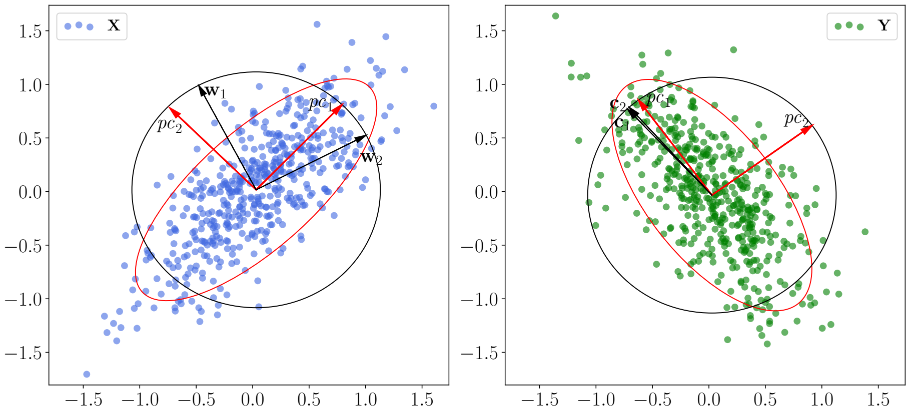
$$\mathbf{x} = \mathbf{P}\mathbf{t} + \mathbf{e}_{\mathbf{x}}$$

$$\mathbf{y} = \mathbf{Q}\mathbf{u} + \mathbf{e}_{\mathbf{y}}$$

$$\text{cov}(\mathbf{t}, \mathbf{u}) \to \max_{\mathbf{P}, \mathbf{Q}}$$

Пример согласованной проекции в скрытое пространство

Исходные переменные $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{\Sigma})$.
Целевые переменные $\mathbf{y}_i$ линейно зависят от pc_2 и не зависят от pc_1.

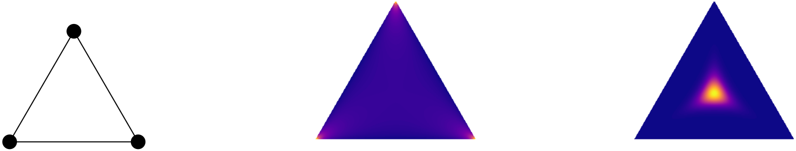


Согласование проекций матриц **X** и **Y** находит оптимальное скрытое представление, отклоняя вектора $\mathbf{w}_k$ и $\mathbf{c}_k$ от направления главных компонент.

Априорное распределение на структуре модели

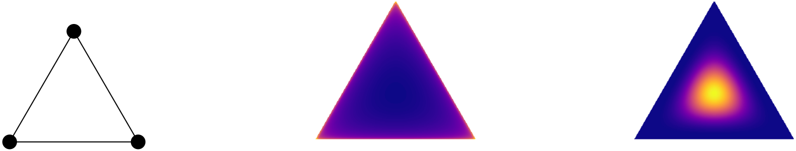
Каждая точка на симплексе задает модель.

Распределение Гумбель-софтмакс: $\boldsymbol{\Gamma} \sim \mathrm{GS}(\mathbf{s}, \lambda_{\text{temp}})$



$\lambda_{\text{temp}} \to 0$ $\qquad$ $\lambda_{\text{temp}} = 0.995$ $\qquad$ $\lambda_{\text{temp}} = 5.0$

Распределение Дирихле: $\boldsymbol{\Gamma} \sim \mathrm{Dir}(\mathbf{s}, \lambda_{\text{temp}})$

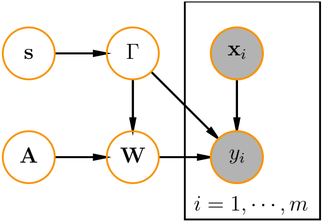


$\lambda_{\text{temp}} \to 0$ $\qquad$ $\lambda_{\text{temp}} = 0.995$ $\qquad$ $\lambda_{\text{temp}} = 5.0$

Байесовский выбор модели

Базовая модель:

- **параметры** модели $\mathbf{w} \sim \mathcal{N}(0, \alpha^{-1})$,
- **гиперпараметры** модели $\mathbf{h} = [\alpha]$.



Предлагаемая модель:

- **параметры** модели $\mathbf{w}_r^{j,k} \sim \mathcal{N}(0, \gamma_r^{j,k}(\mathbf{A}_r^{j,k})^{-1})$, $\mathbf{A}_r^{j,k}$ — диагональная матрица параметров, соответствующих базовых функций $\mathbf{g}_r^{j,k}$, $(\mathbf{A}_r^{j,k})^{-1} \sim \text{inv-gamma}(\lambda_1, \lambda_2)$,
- **структурные параметры** модели $\mathbf{\Gamma} = \{\gamma^{j,k}, (j,k) \in E\}$, $\gamma^{j,k} \sim \text{GS}(\mathbf{s}^{j,k}, \lambda_{\text{temp}})$,
- **гиперпараметры** модели $\mathbf{h} = [\text{diag}(\mathbf{A}), \mathbf{s}]$,
- **метапараметры** $\lambda_1, \lambda_2, \lambda_{\text{temp}}$.

Оптимальные факторы $q^\star(\boldsymbol{\theta}), q^\star(\mathbf{m}, \mathbf{V}, \boldsymbol{\alpha})$ имеют вид

$$\ln q^\star = \mathsf{E}\big[\ln p(\mathbf{Z}, \boldsymbol{\theta}, \mathbf{m}, \mathbf{V}, \boldsymbol{\alpha})\big] + \text{const}$$

и не вычисляются аналитически, так как правдоподобие L в модели $p(\mathbf{Z}, \boldsymbol{\theta}, \mathbf{m}, \mathbf{V}, \boldsymbol{\alpha})$ содержит сумму экспонент $g(\mathbf{s}_n)$ в знаменателе,

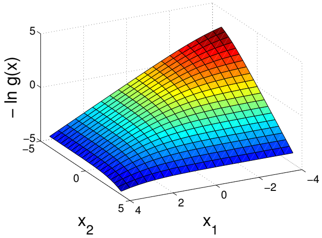
$$L(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\alpha}) = \prod_{n=1}^{N} \prod_{k=1}^{K_h} \left[\frac{\exp(s_{n,k})}{g(\mathbf{s}_n)}\right]^{z_{nk}}, \quad g(\mathbf{s}_n) = \sum_{k=1}^{K_h} \exp(s_{n,k}).$$

$\ln \frac{1}{g(\mathbf{x})}$ – вогнутая функция. Касательная плоскость к ней через точку $\boldsymbol{\xi}$:

$$\zeta(\mathbf{x}, \boldsymbol{\xi}) = -\ln\big(g(\boldsymbol{\xi})\big) - \nabla \ln\big(g(\boldsymbol{\xi})\big)^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\xi}).$$

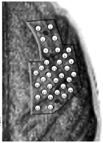
Верхняя оценка правдоподобия L

$$\frac{1}{g(\boldsymbol{\xi})} \exp\left(s_{n,k} + \sum_{k'=1}^{K_h} \frac{\exp(\xi_{k'})}{g(\boldsymbol{\xi})}(\xi_{k'} - s_{n,k'})\right).$$



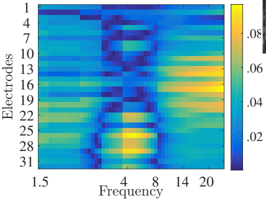
Анализ качества предлагаемого метода выбора признаков

Решена задача прогнозирования многомерных временных рядов $\mathbf{y}(t) \in \mathbb{R}^3$ координат конечности по интервалам $\mathbf{s}(t - \Delta t)$ многомерных временных рядов $\mathbf{s}(t) \in \mathbb{R}^{N_{\mathbf{ch}}}$ многоканальных электрокортикограмм.



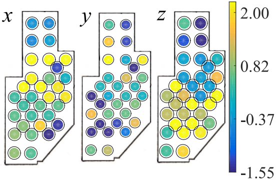
Признаки:

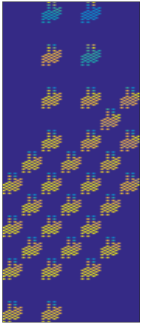
$$\underline{\mathbf{X}}_m \in \mathbb{R}^{F \times N_{\mathbf{ch}}}, \quad \underline{\mathbf{X}}_{mjn} = \begin{cases} s_n(t_m + \tau), \; j = 1, \\ W_{mjn}, \; j = 2, \dots, F+1, \end{cases}$$



Прогноз в точке t_m:

$$\hat{\mathbf{y}}_m = \operatorname{vec}\left(\underline{\mathbf{X}}_m\right)^{\mathsf{T}} \hat{\mathbf{w}}.$$





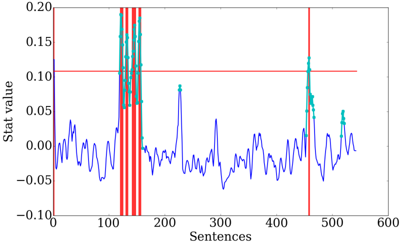
Качество прогнозирования:
коэффициент корреляции между $\hat{\mathbf{Y}}$ и $\mathbf{Y}$,

$$\operatorname{corr}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\operatorname{cov}(\hat{\mathbf{y}}, \mathbf{y})}{\sqrt{\operatorname{cov}(\hat{\mathbf{y}}, \hat{\mathbf{y}})\operatorname{cov}(\mathbf{y}, \mathbf{y})}}.$$

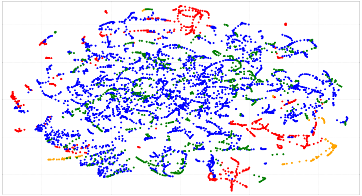
масштабированная ошибка MSE,

$$\operatorname{sMSE}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\sum_{m=1}^{M} ||\hat{\mathbf{y}}_m - \mathbf{y}_m||_2}{\sum_{m=1}^{M} ||\bar{\mathbf{y}} - \mathbf{y}_m||_2}.$$

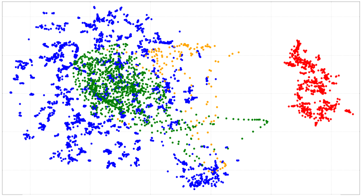
Снижение размерности с сохранением локальной структуры близости



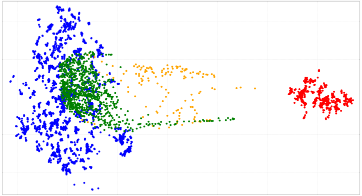
Построение признакового пространства для ряда **s**



Исходный t-SNE



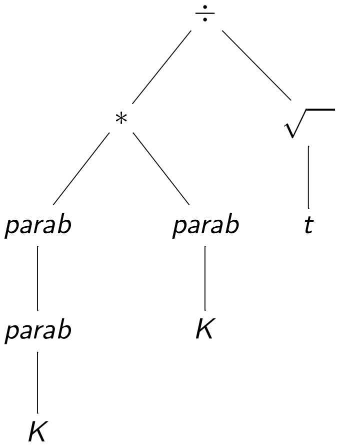
Модифицированный t-SNE, $\mu = 10$

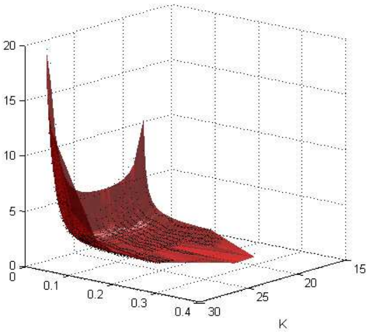


Модифицированный t-SNE, $\mu = 100$

Результаты для опциона CLZ

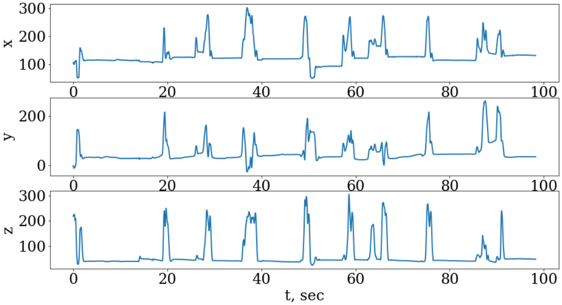
$$\sigma = \frac{(w_1 K + w_2)(w_1 K^2 + w_2 K + w_3)^2}{\sqrt{t}}$$



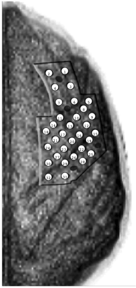


Декодируемые сигналы электрокортикограммы

- Сигналы $\mathbf{s}(t) \in \mathbb{R}^{N_{\mathbf{ch}}}$. 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



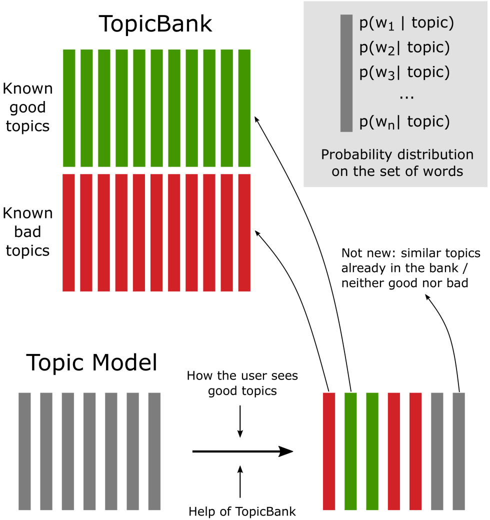
Пространственное расположение электродов

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 $p(\Phi, \Theta \mid D)$.

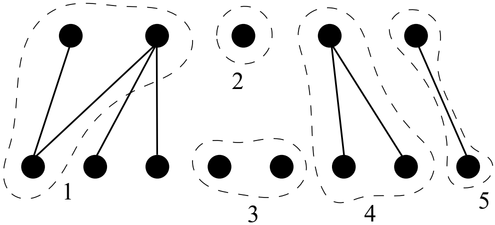


Построение банка тем

Аналогично построению двухуровневой иерархической тематической модели:

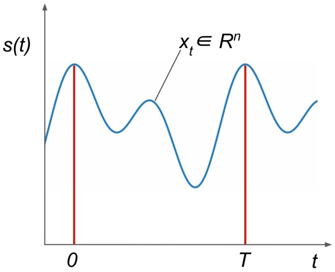
$$\underbrace{p(w \mid t)}_{\varphi_{wt}^{parent}} = \sum_{s \in S} \underbrace{p(w \mid s)}_{\varphi_{ws}^{child}} \underbrace{p(s \mid t)}_{\psi_{st}} \qquad \text{Hierarchy}$$

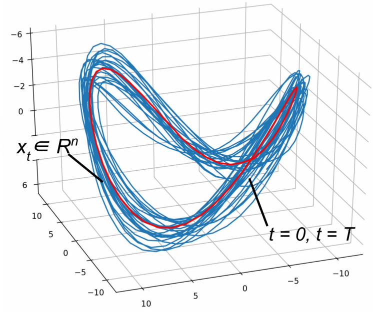
$$\underbrace{p(w \mid t)}_{\varphi_{wt}^{\mathbf{bank}}} = \sum_{s \in S} \underbrace{p(w \mid s)}_{\varphi_{ws}^{\mathbf{new}}} \underbrace{p(s \mid t)}_{\psi_{st}} \qquad \text{TopicBank}$$



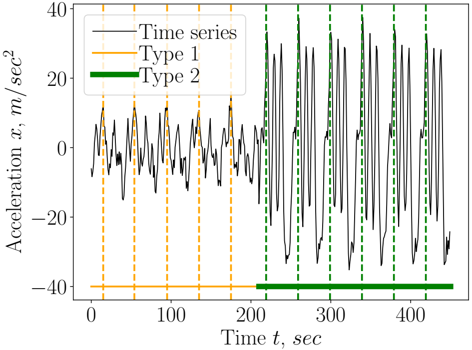
№	Hierarchy	TopicBank
1	ok	no
2	ok	ok
3	no	ok
4	ok	maybe
5	ok	maybe

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.

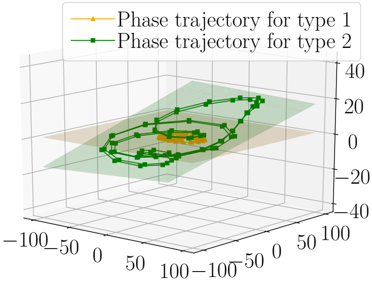




Сегмент — последовательность точек временного ряда, которая относится к одному характерному физическому действию человека: шаг, прыжок.
Цепь — последовательность сегментов, которые образуют квазипериодическую последовательность точек.



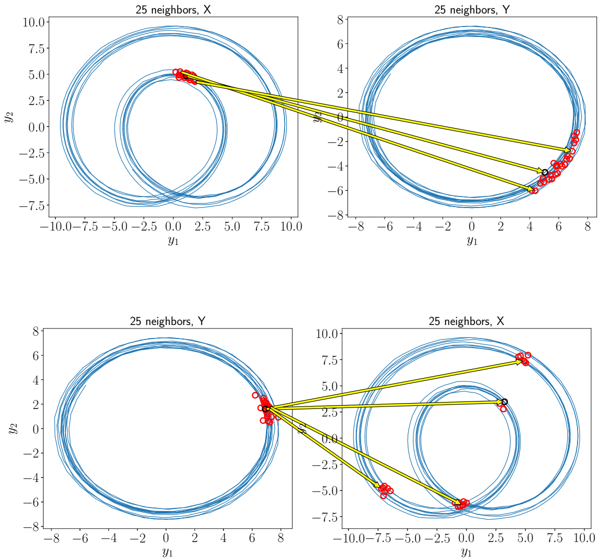
(a)



(b)

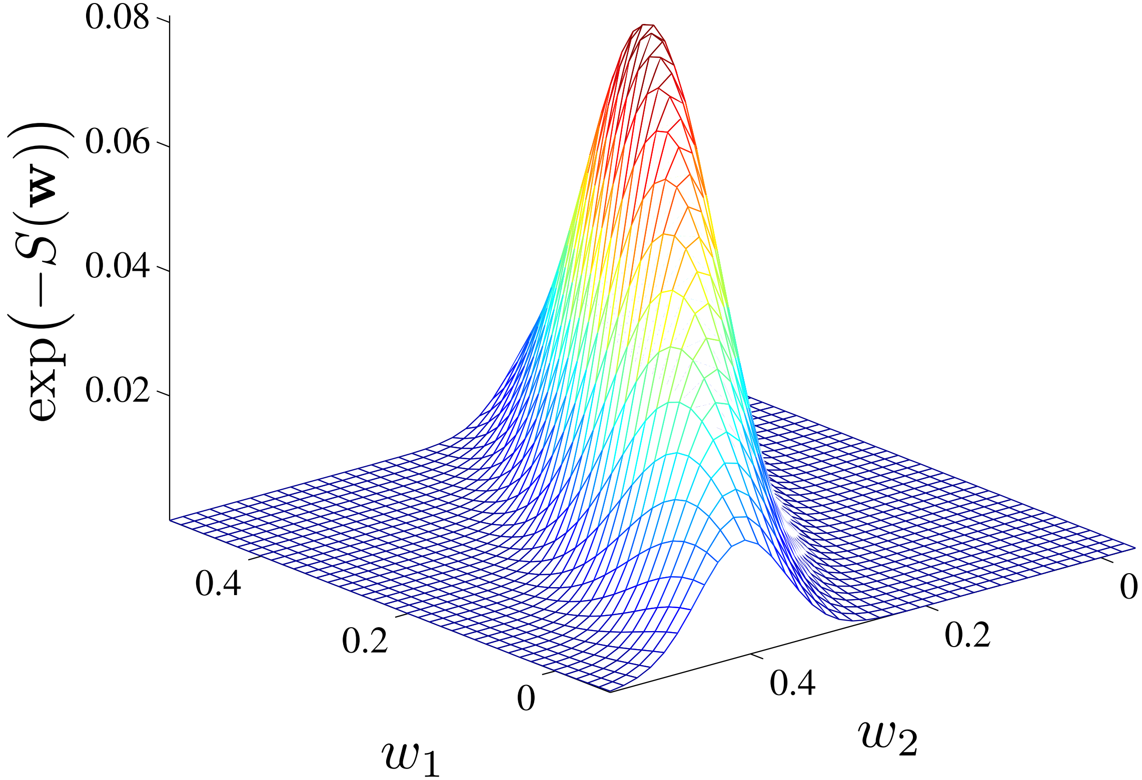
a) временной ряда разбитый на сегменты; b) проекции на плоскость фазовых траекторий временного ряда, которые относятся к Type 1 и Type 2.

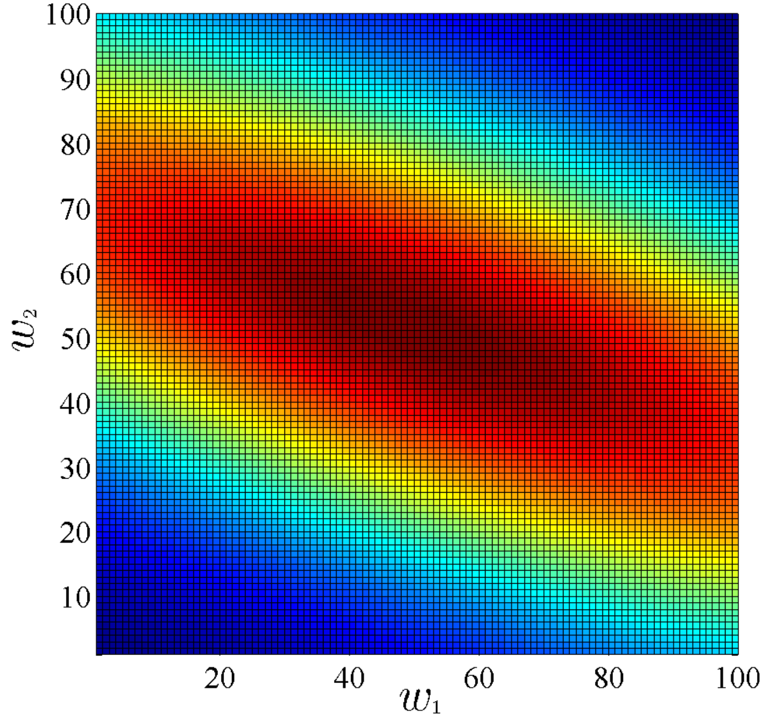
Ближайшие соседи на фазовых траекториях



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)

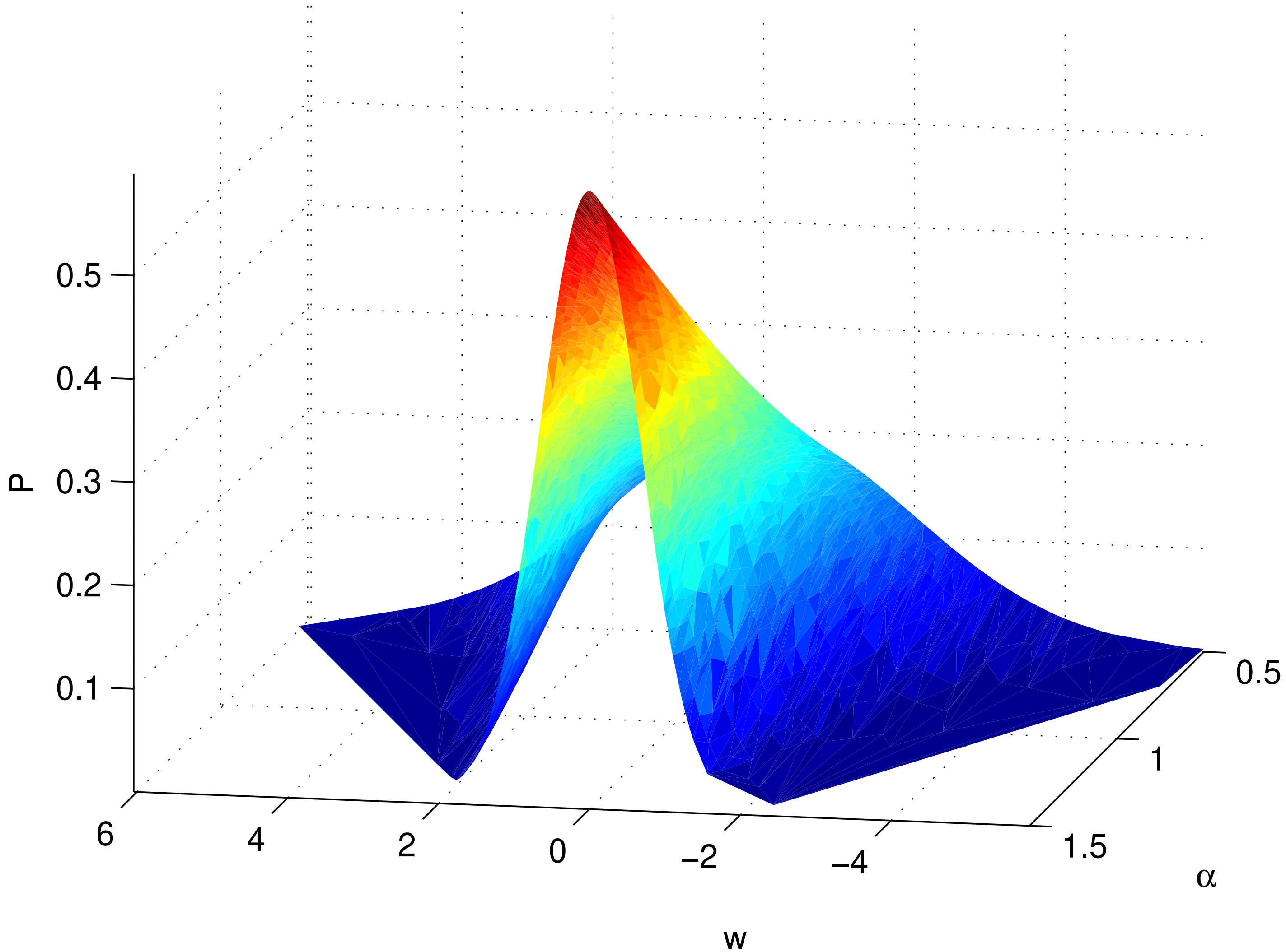
Write the error function given hyperparameters $\mathbf{A},\mathbf{B}$

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

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

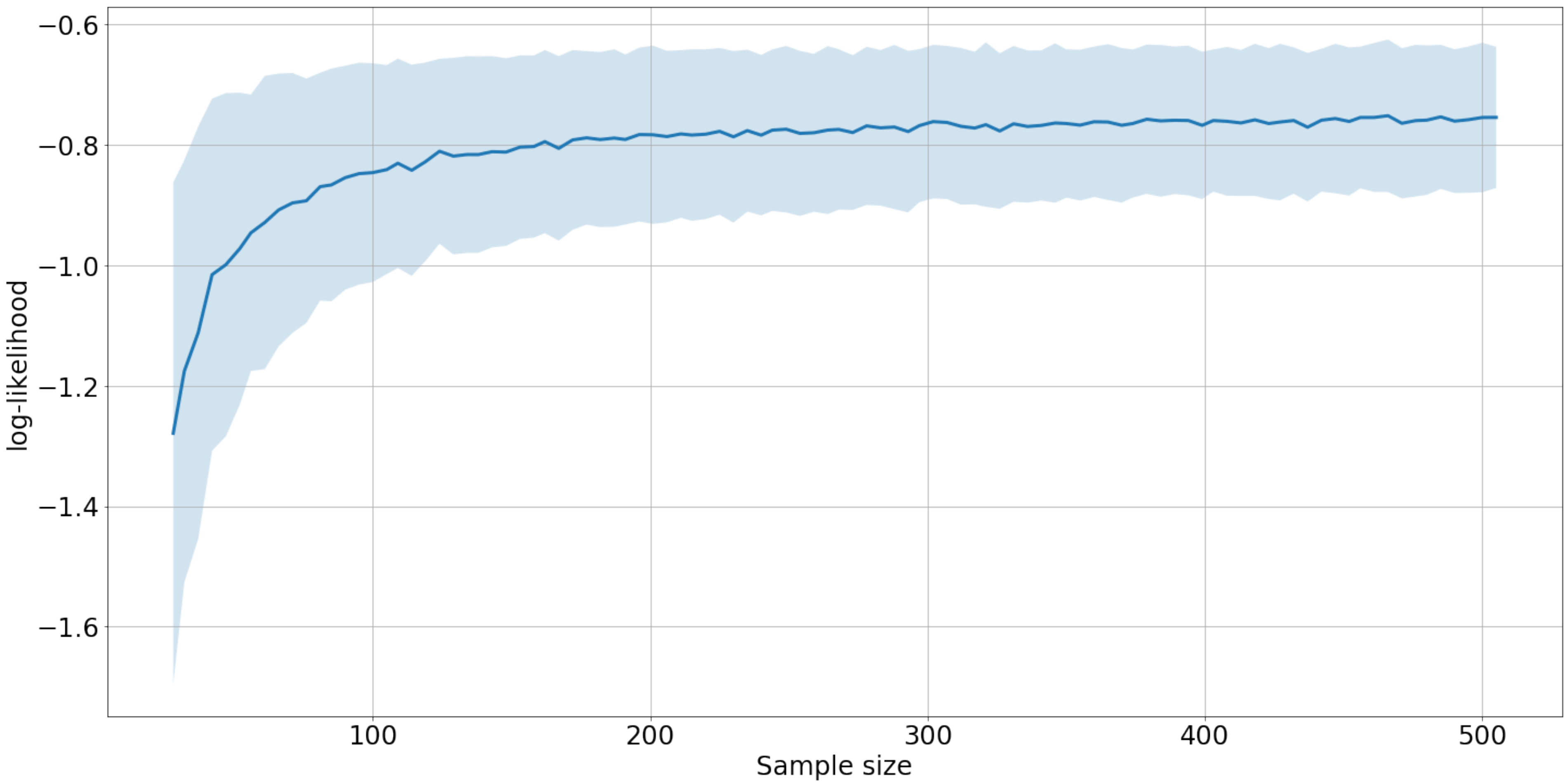
Evidence of the model

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

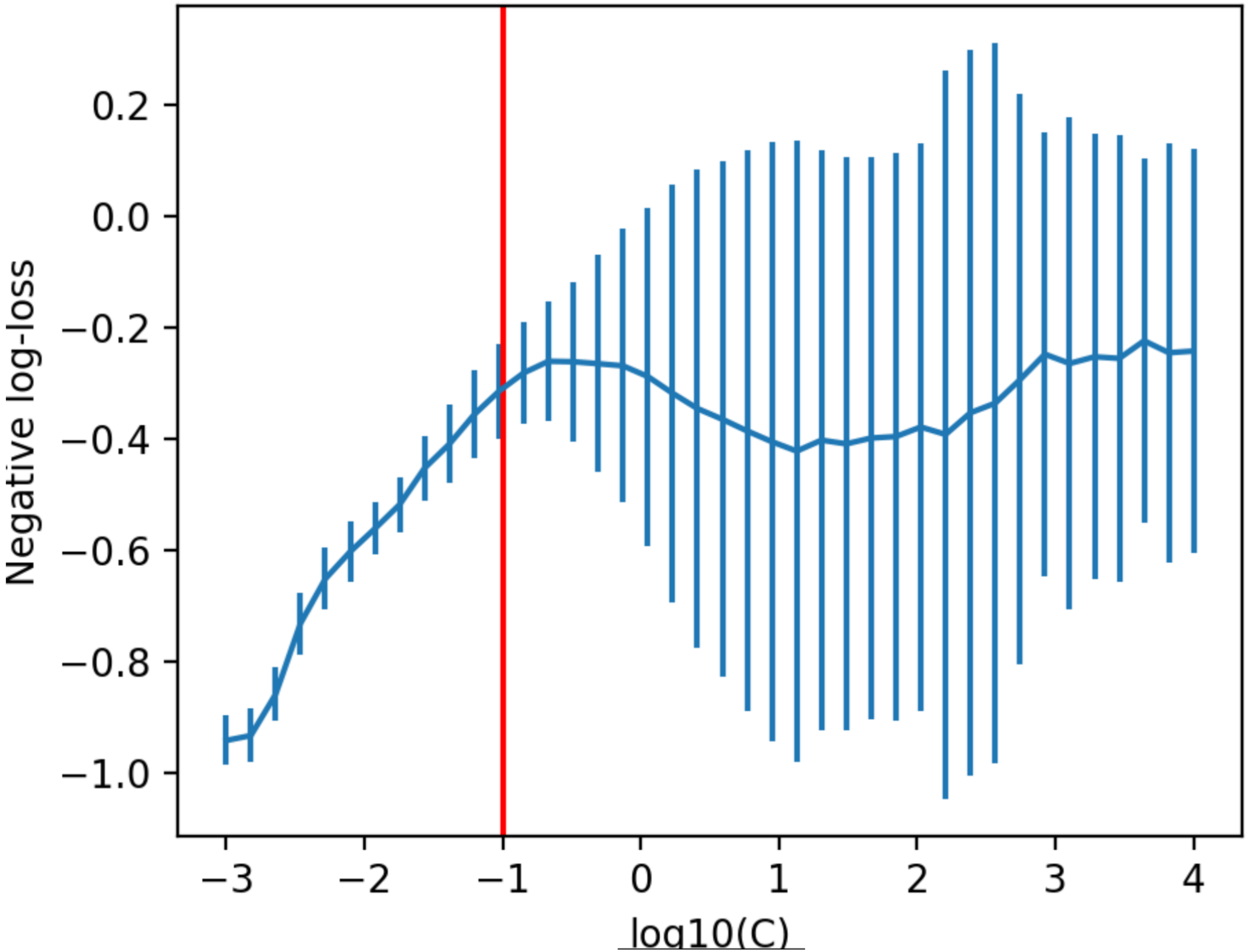


Parameters $\mathbf{w}$, variance α^{-2}, and $p(\mathbf{w}|\mathfrak{D}, \alpha)$ is the evidence.

– Error and its variance for a reinforced sample set



Variance of error increasing over model complexity

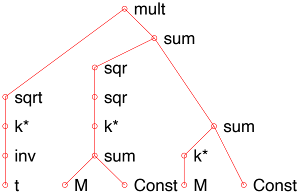


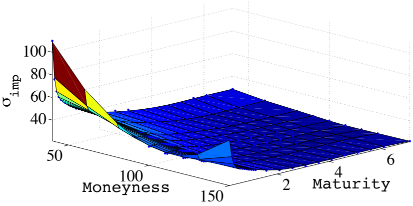
Complexity, number of model parameters

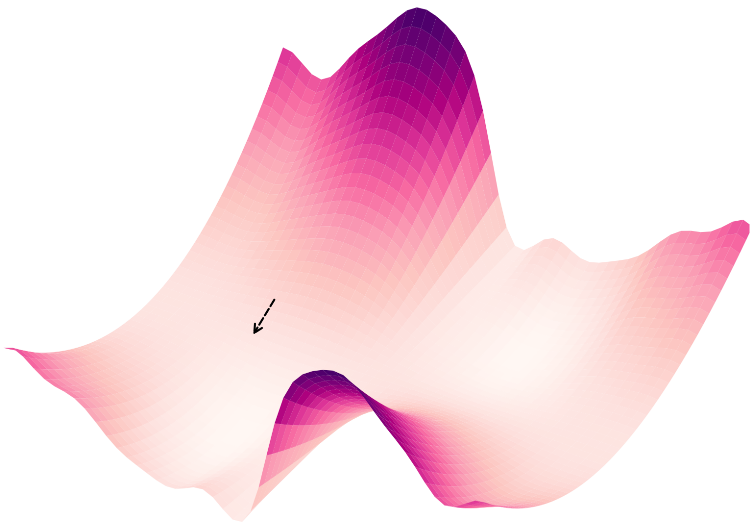
Resulting models

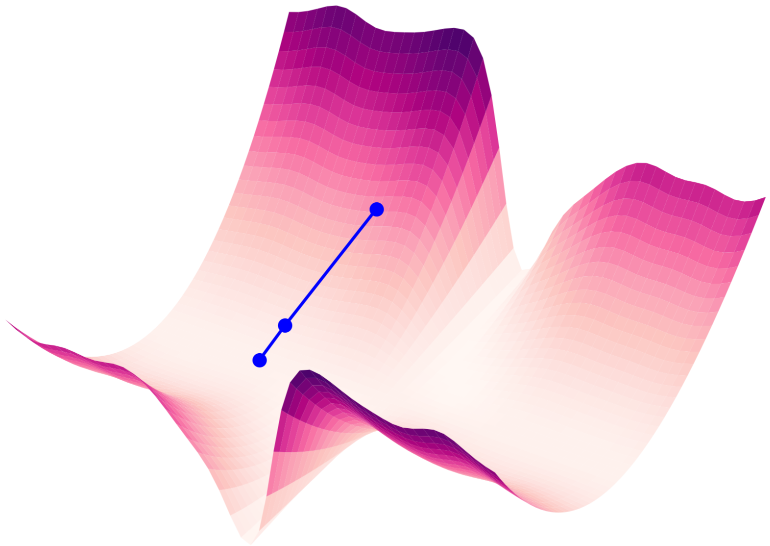
Resulting model

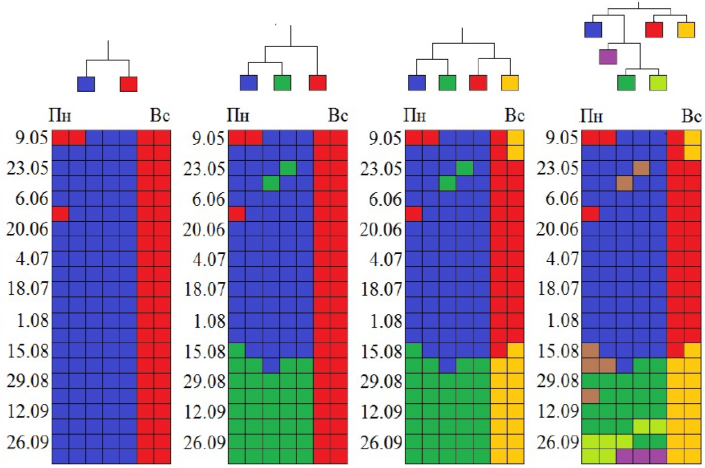
$$\sigma_{imp} = \frac{(k_2 M + k_3)^4 + k_4 M + c}{\sqrt{k_1 t}}$$

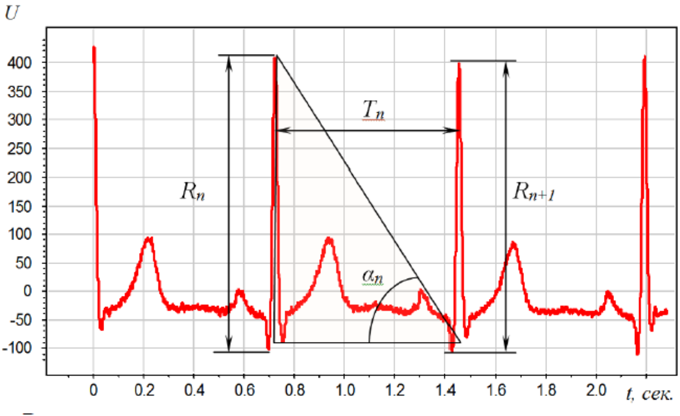




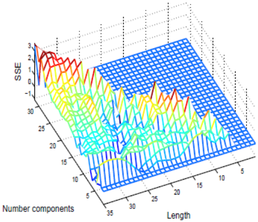


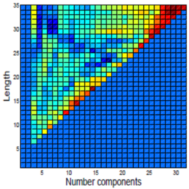



Пн
Вс
9.05
23.05
6.06
20.06
4.07
18.07
1.08
15.08
29.08
12.09
26.09

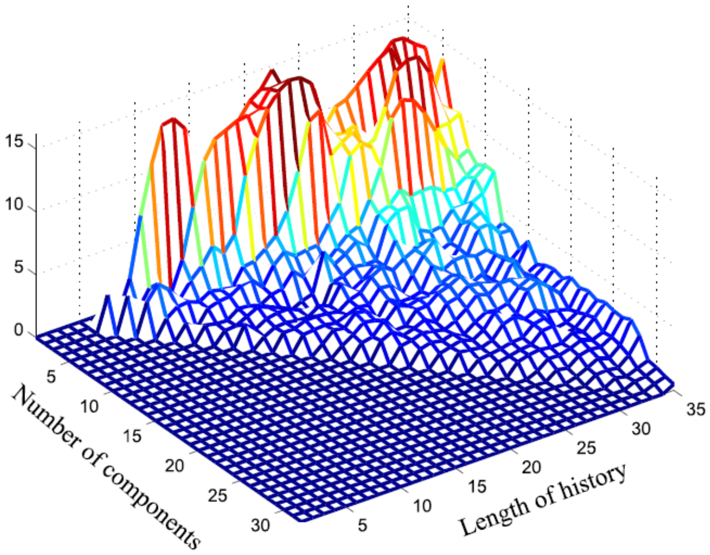

U
400
350
300
250
200
150
100
50
0
-50
-100
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
t, сек.
T_n
R_n
R_{n+1}
α_n

jmlda01

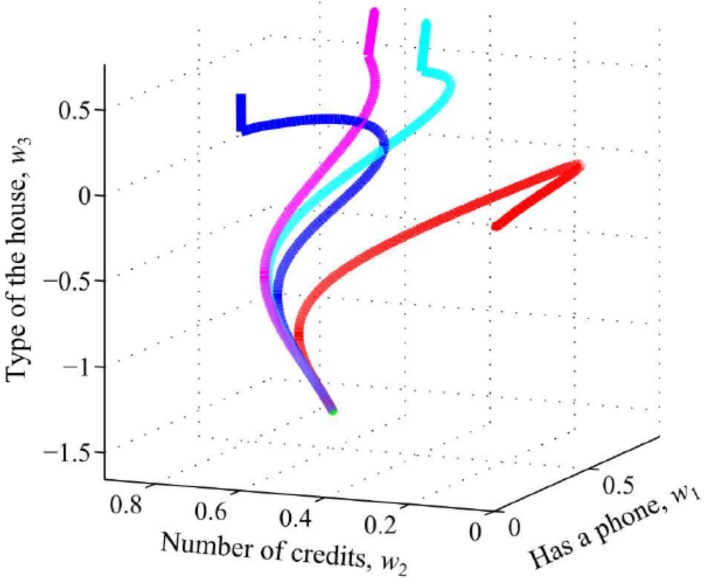


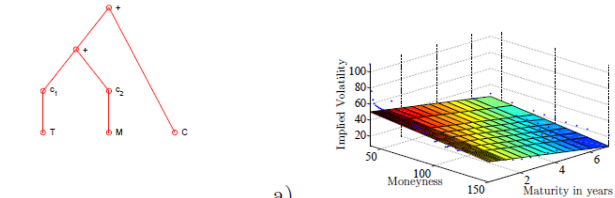


jmlda01

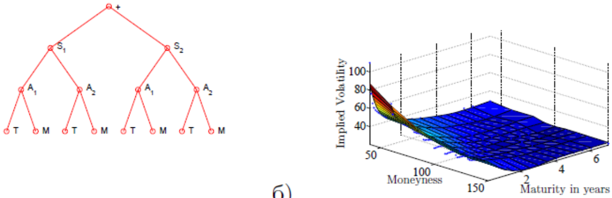


15
10
5
0
5
10
15
20
25
30
Number of components
5
10
15
20
25
30
35
Length of history

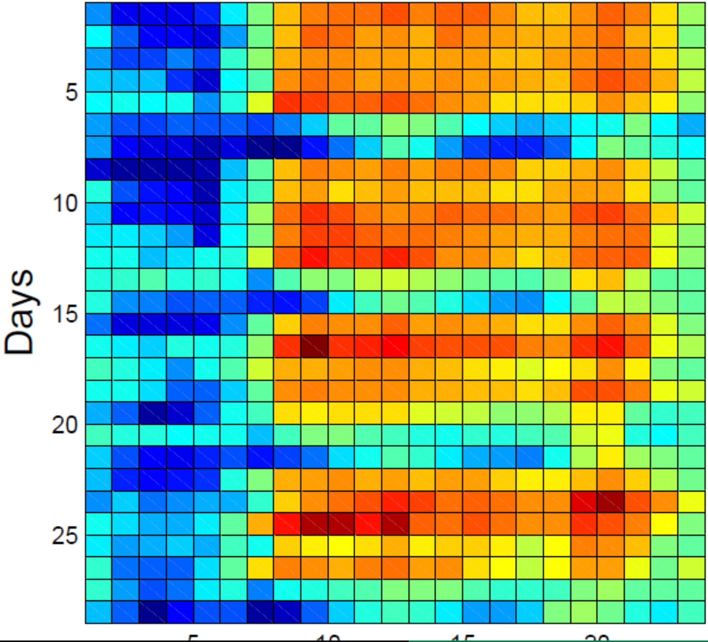

Type of the house, w_3
0.5
0
−0.5
−1
−1.5
0.8
0.6
0.4
0.2
0
Number of credits, w_2
0
0.5
Has a phone, w_1

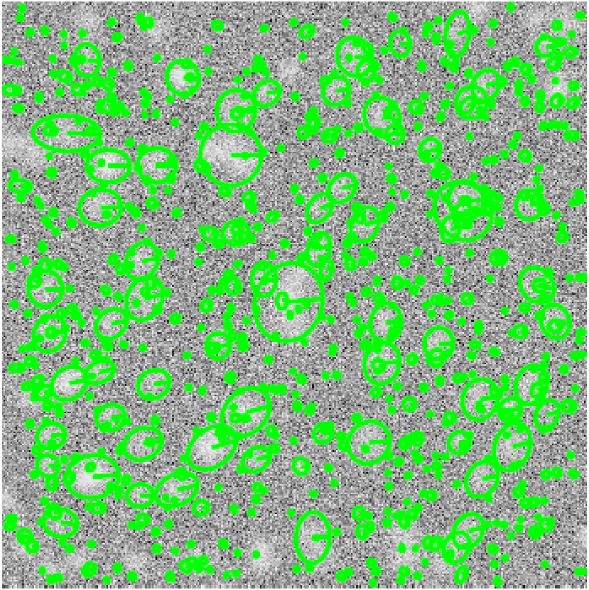

+
+
c1
c2
T
M
C
Implied Volatility
100
80
60
40
20
50
100
150
Moneyness
2
4
6
Maturity in years

a)

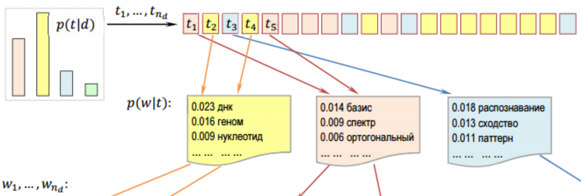

+
S1
S2
A1
A2
A1
A2
T
M
T
M
T
M
T
M
Implied Volatility
100
80
60
40
50
100
150
Moneyness
2
4
6
Maturity in years

б)


Days
5
10
15
20
25

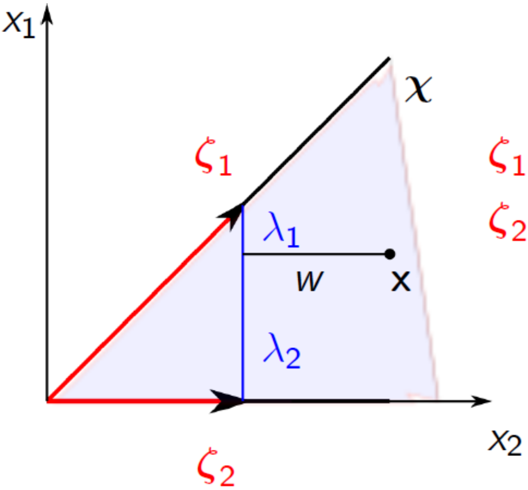


Тематическое моделирование и матричное разложение



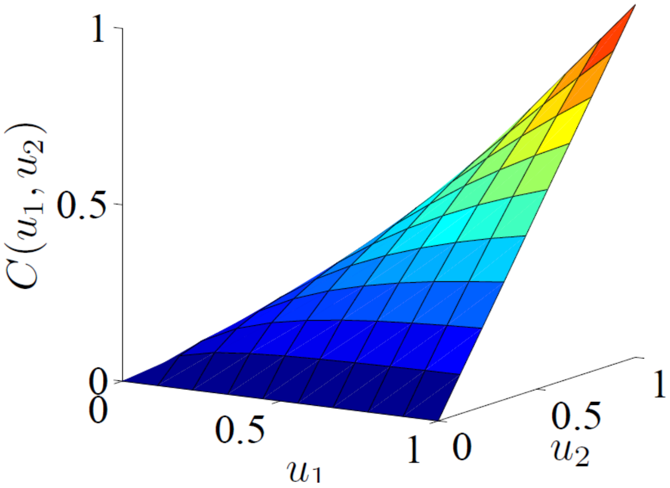
Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

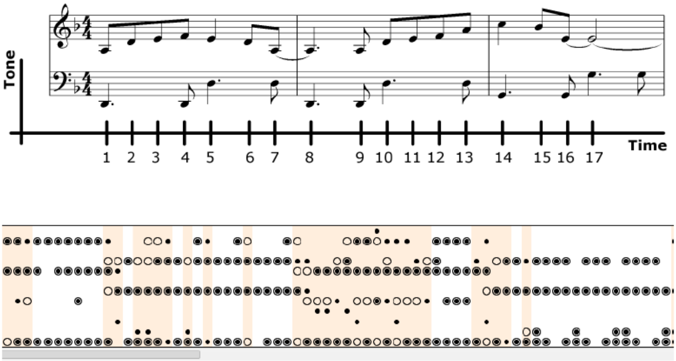
Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization // International Symposium On Learning And Data Sciences 2015.



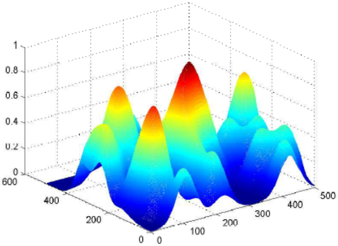
$\zeta_1 = [1; 1]$,

$\zeta_2 = [1; 0]$

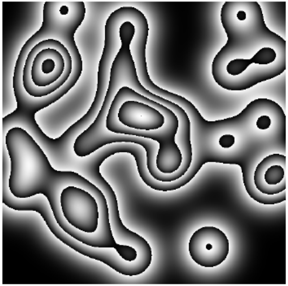

$C(u_1, u_2)$
1
0.5
0
0
0.5
1
u_1
0
0.5
1
u_2



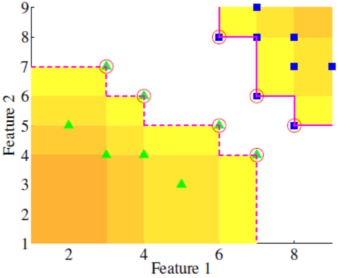
Пустые круги — истинные полутона, точки — предсказанные, ошибки подсвечены, горизонтальная ось — время.



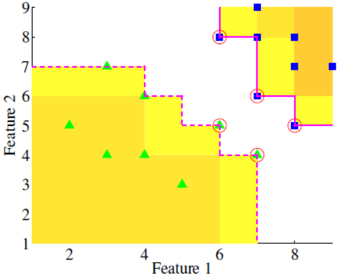
(а) Высота



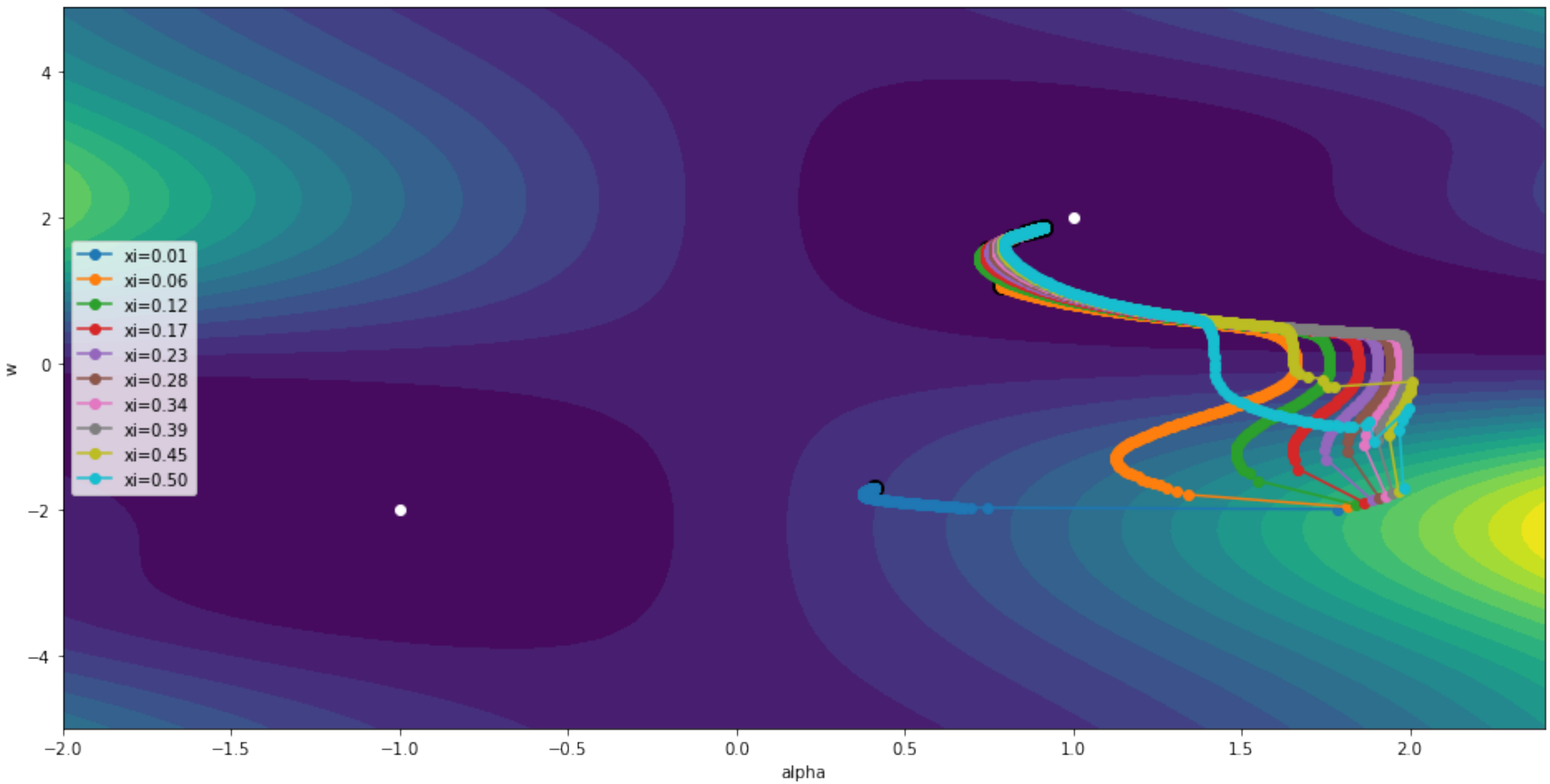
(б) Фазовая составляющая без учета шума



Feature 2
Feature 1



Feature 2
Feature 1


xi=0.01
xi=0.06
xi=0.12
xi=0.17
xi=0.23
xi=0.28
xi=0.34
xi=0.39
xi=0.45
xi=0.50
w
alpha
4
2
0
−2
−4
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0