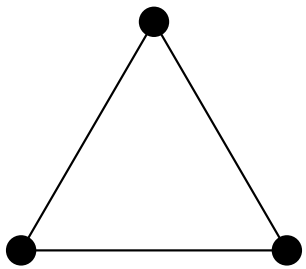
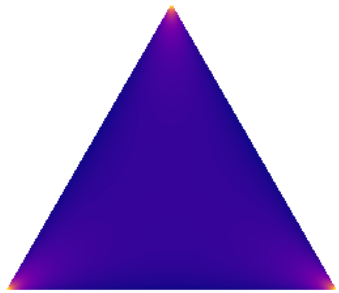
Априорное распределение на структуре модели

Каждая точка на симплексе задает модель.

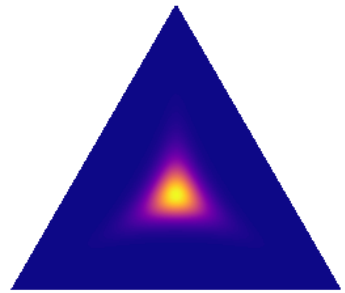
Распределение Гумбель-софтмакс: $\mathbf{\Gamma} \sim \mathrm{GS}(\mathbf{s}, \lambda_{\text{temp}})$



$\lambda_{\text{temp}} \to 0$

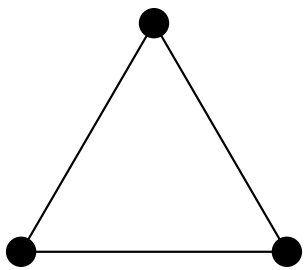


$\lambda_{\text{temp}} = 0.995$

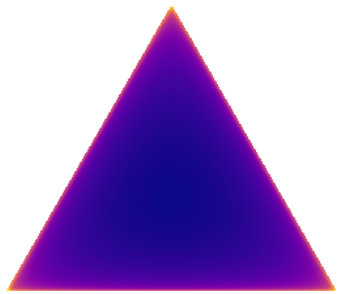


$\lambda_{\text{temp}} = 5.0$

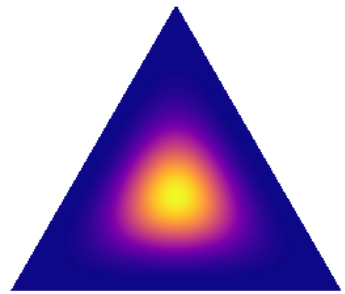
Распределение Дирихле: $\mathbf{\Gamma} \sim \mathrm{Dir}(\mathbf{s}, \lambda_{\text{temp}})$



$\lambda_{\text{temp}} \to 0$



$\lambda_{\text{temp}} = 0.995$

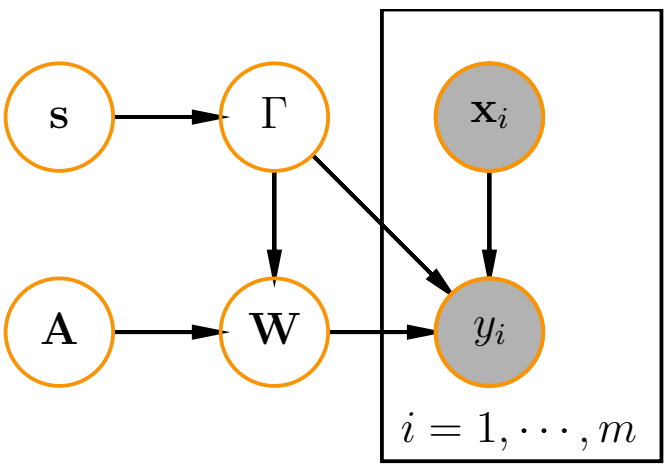


$\lambda_{\text{temp}} = 5.0$

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 $\mathbf{\Gamma} = \{\boldsymbol{\gamma}^{j,k}, (j,k) \in E\}$, $\boldsymbol{\gamma}^{j,k} \sim \text{GS}(\mathbf{s}^{j,k}, \lambda_{\text{temp}})$,
- **гиперпараметры** модели $\mathbf{h} = [\text{diag}(\mathbf{A}), \mathbf{s}]$,
- **метапараметры** $\lambda_1, \lambda_2, \lambda_{\text{temp}}$.

Верхняя оценка правдоподобия модели

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

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

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,

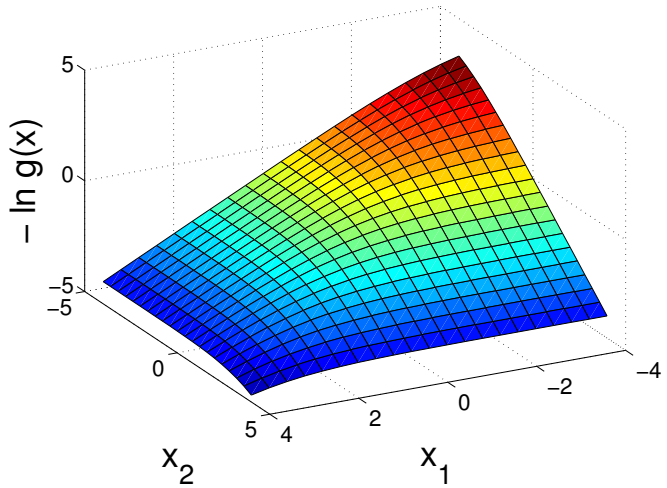
$$L(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\alpha}) = \prod_{n=1}^{N} \prod_{k=1}^{K_h} \left[\frac{\exp(s_{n,k})}{g(\mathbf{s}_n)}\right]^{z_{nk}}, \quad g(\mathbf{s}_n) = \sum_{k=1}^{K_h} \exp(s_{n,k}).$$

$\ln \frac{1}{g(\mathbf{x})}$ – 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:

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

Прогноз в точке t_m:

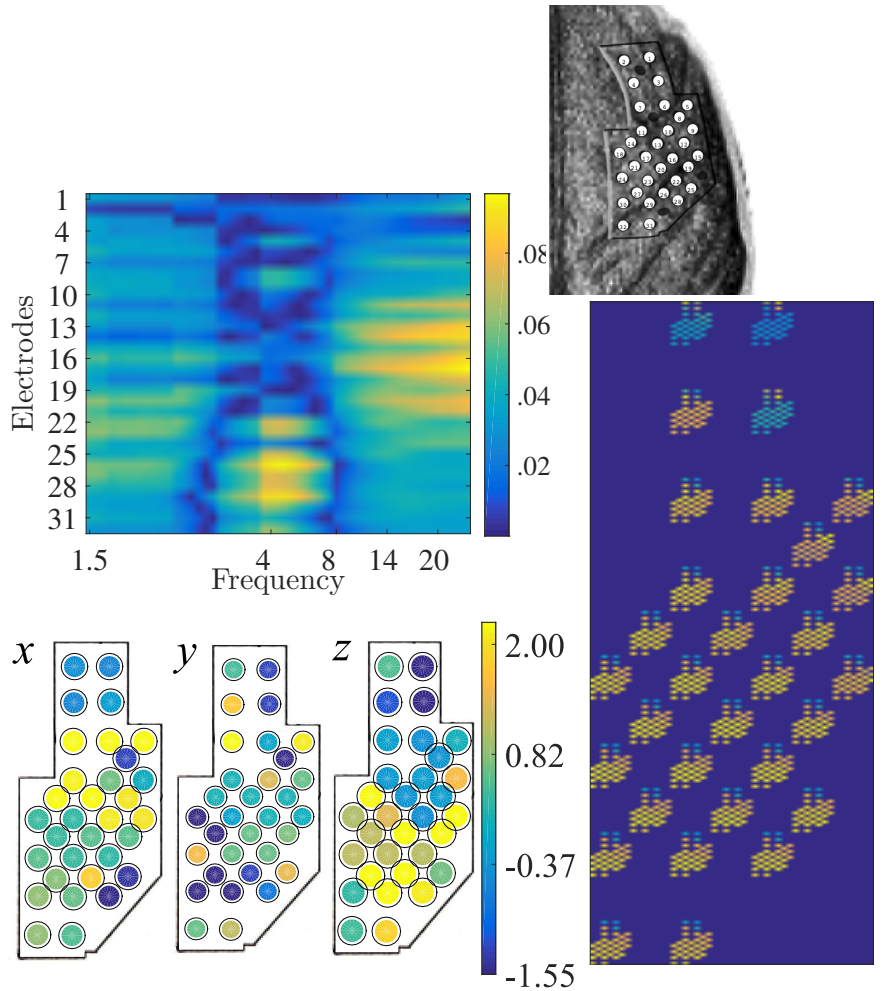
$$\hat{\mathbf{y}}_m = \mathrm{vec}\left(\underline{\mathbf{X}}_m\right)^{\mathsf{T}} \hat{\mathbf{w}}.$$

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

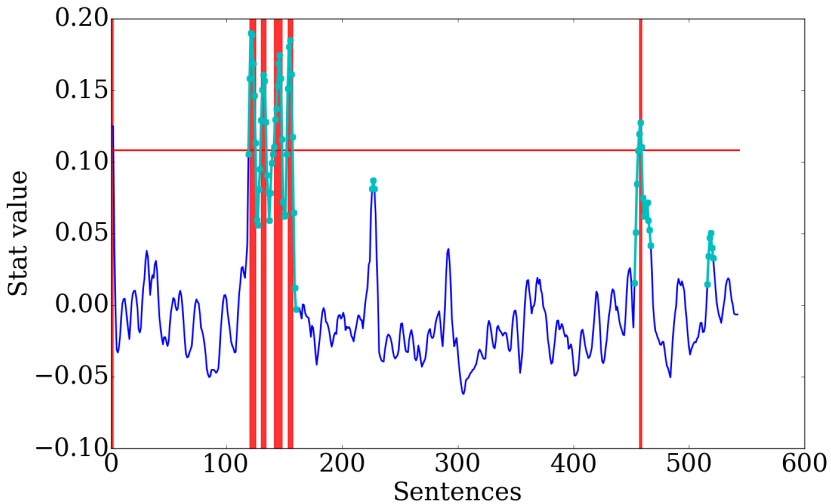
$$\mathrm{corr}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\mathrm{cov}(\hat{\mathbf{y}}, \mathbf{y})}{\sqrt{\mathrm{cov}(\hat{\mathbf{y}}, \hat{\mathbf{y}})\mathrm{cov}(\mathbf{y}, \mathbf{y})}}.$$

масштабированная ошибка MSE,

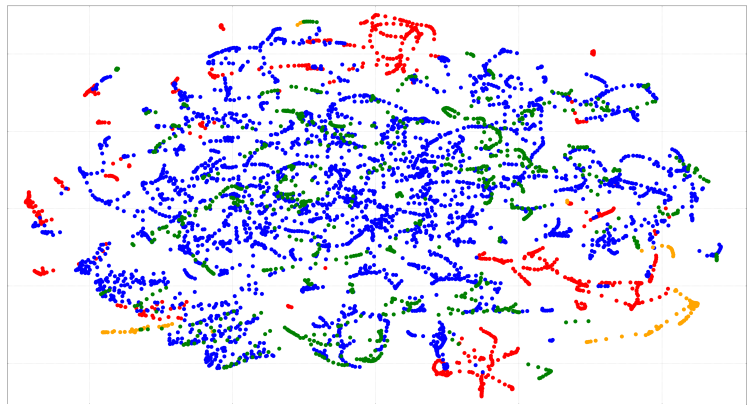
$$\mathrm{sMSE}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\sum_{m=1}^{M} ||\hat{\mathbf{y}}_m - \mathbf{y}_m||_2}{\sum_{m=1}^{M} ||\bar{\mathbf{y}} - \mathbf{y}_m||_2}.$$



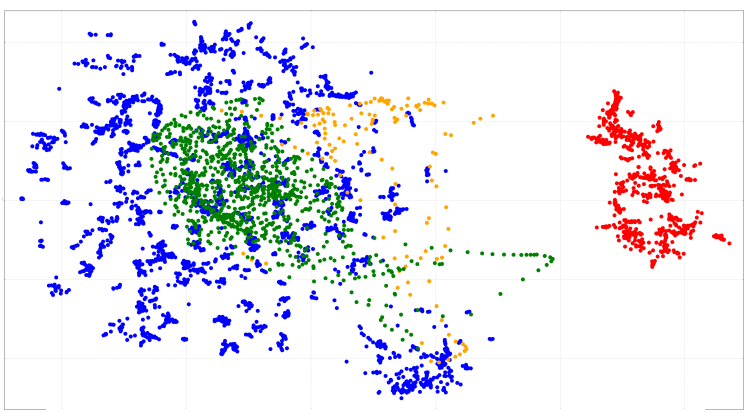
Снижение размерности с сохранением локальной структуры близости



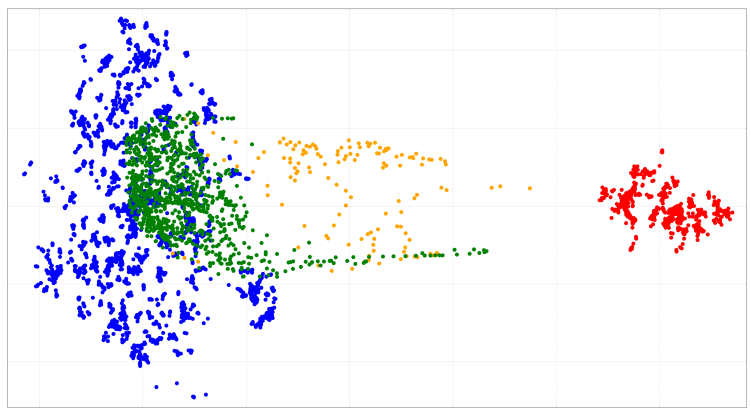
Построение признакового пространства для ряда **s**



Исходный t-SNE



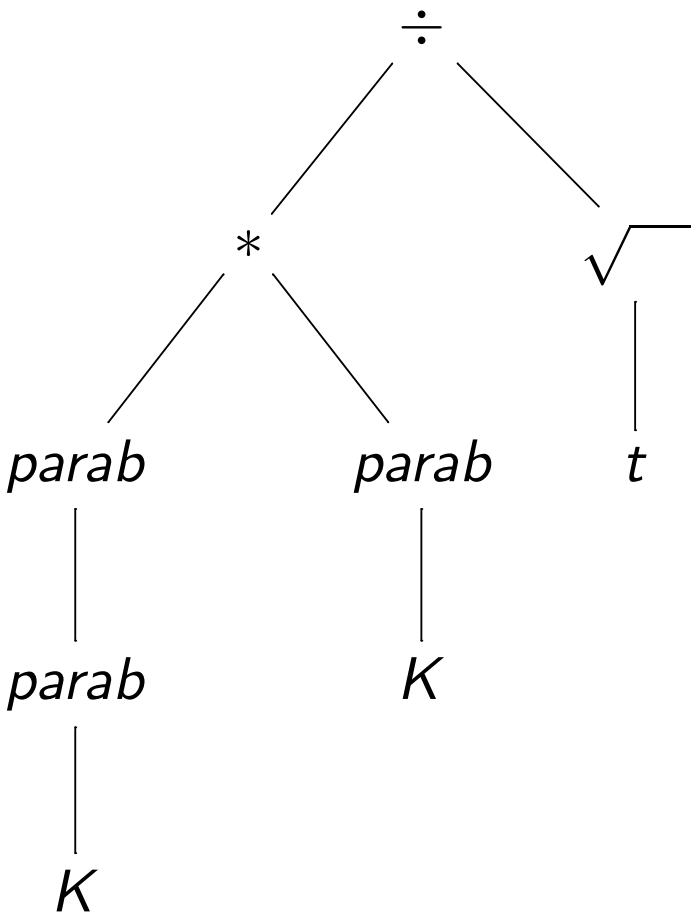
Модифицированный t-SNE, $\mu = 10$

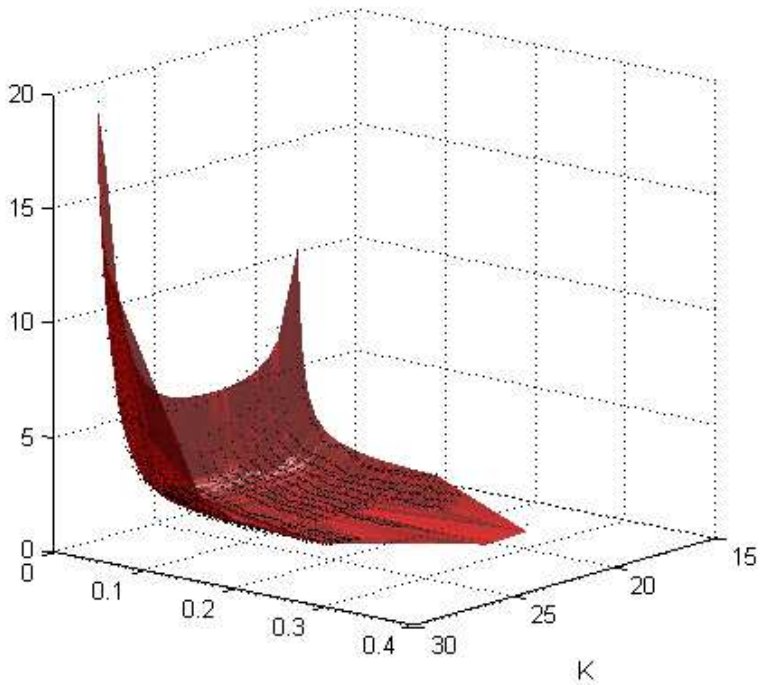


Модифицированный t-SNE, $\mu = 100$

Результаты для опциона CLZ

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

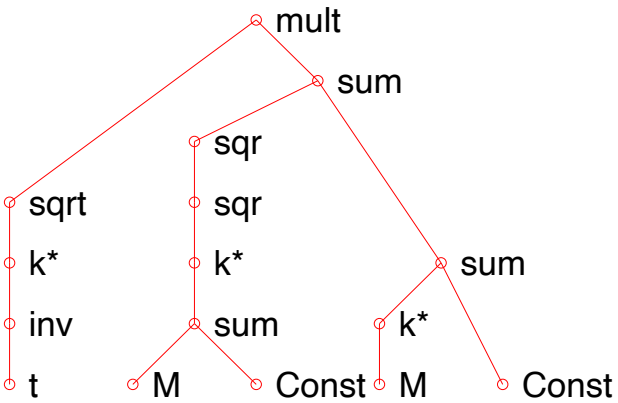


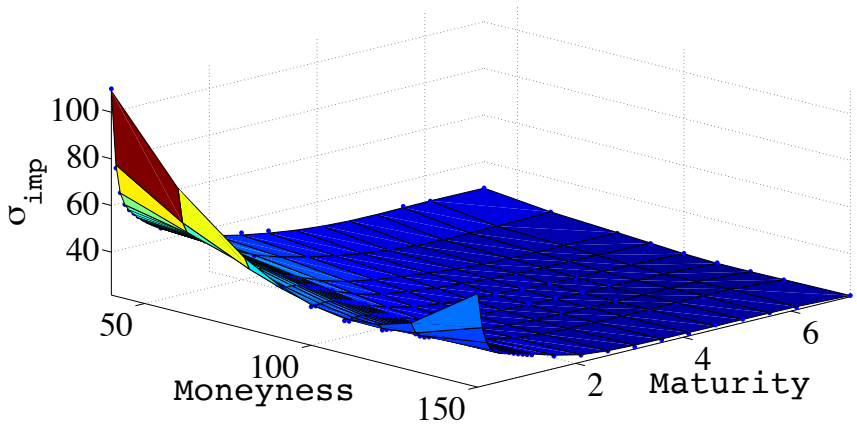


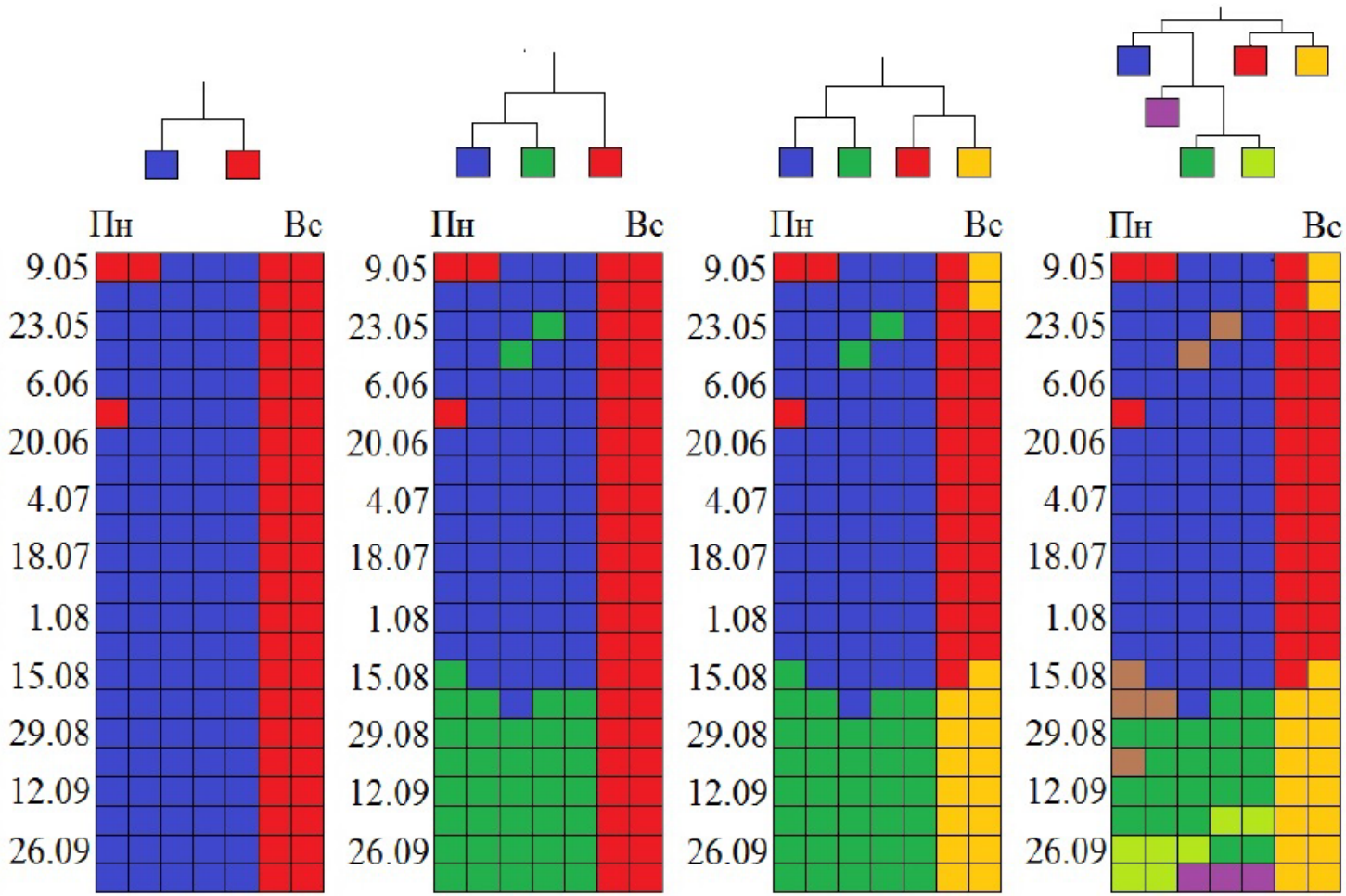
Resulting models

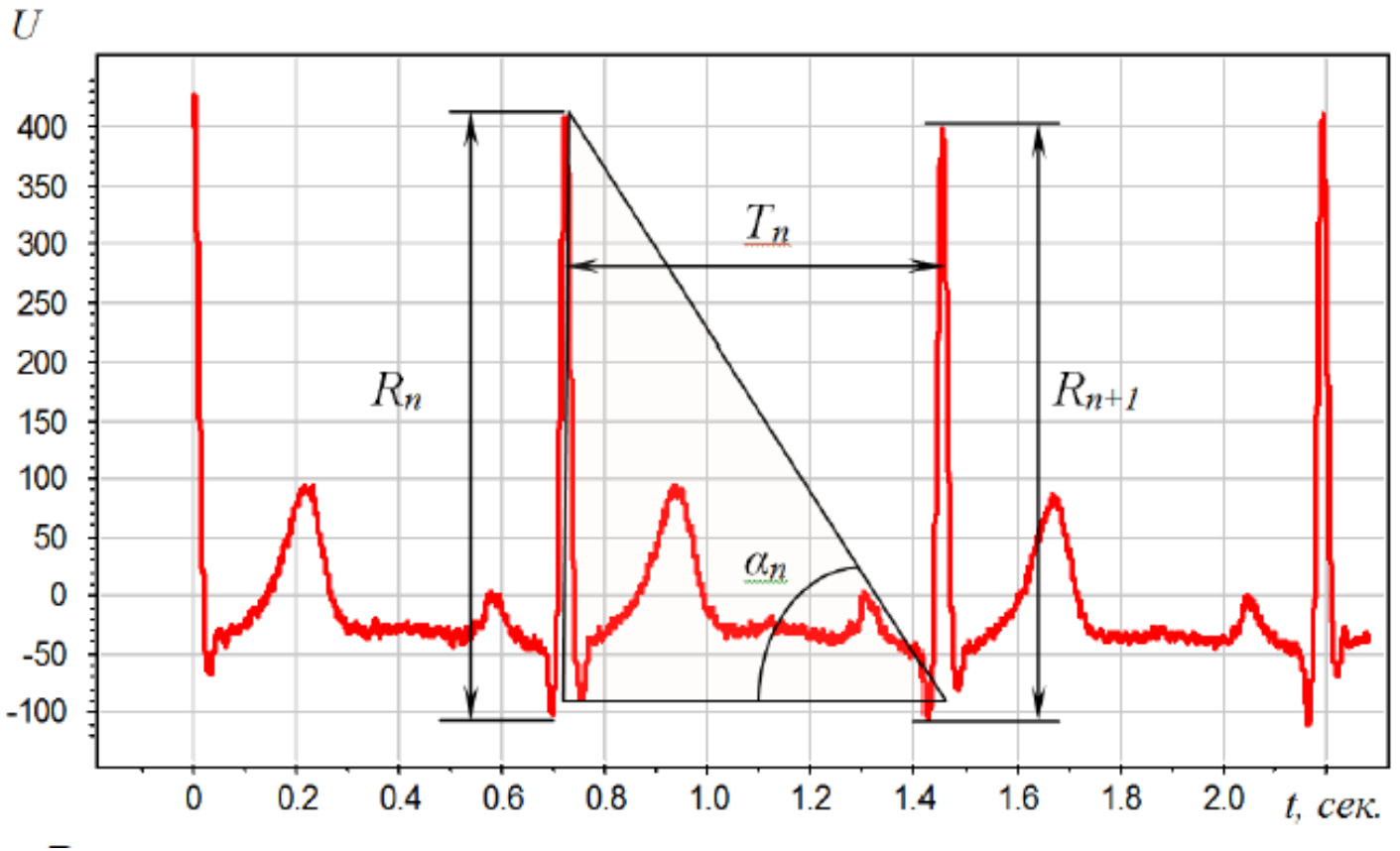
Resulting model

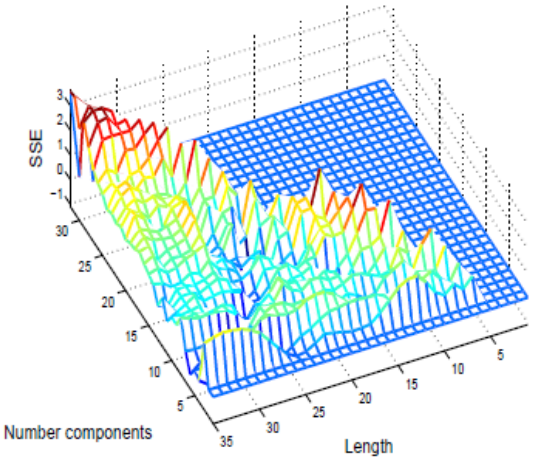
$$\sigma_{imp} = \frac{(k_2 M + k_3)^4 + k_4 M + c}{\sqrt{k_1 t}}$$

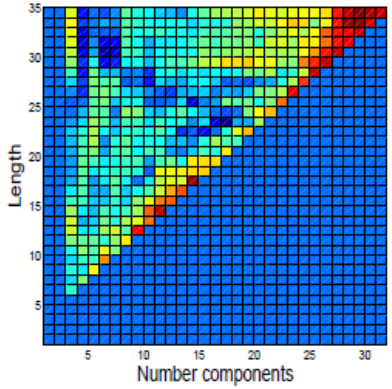


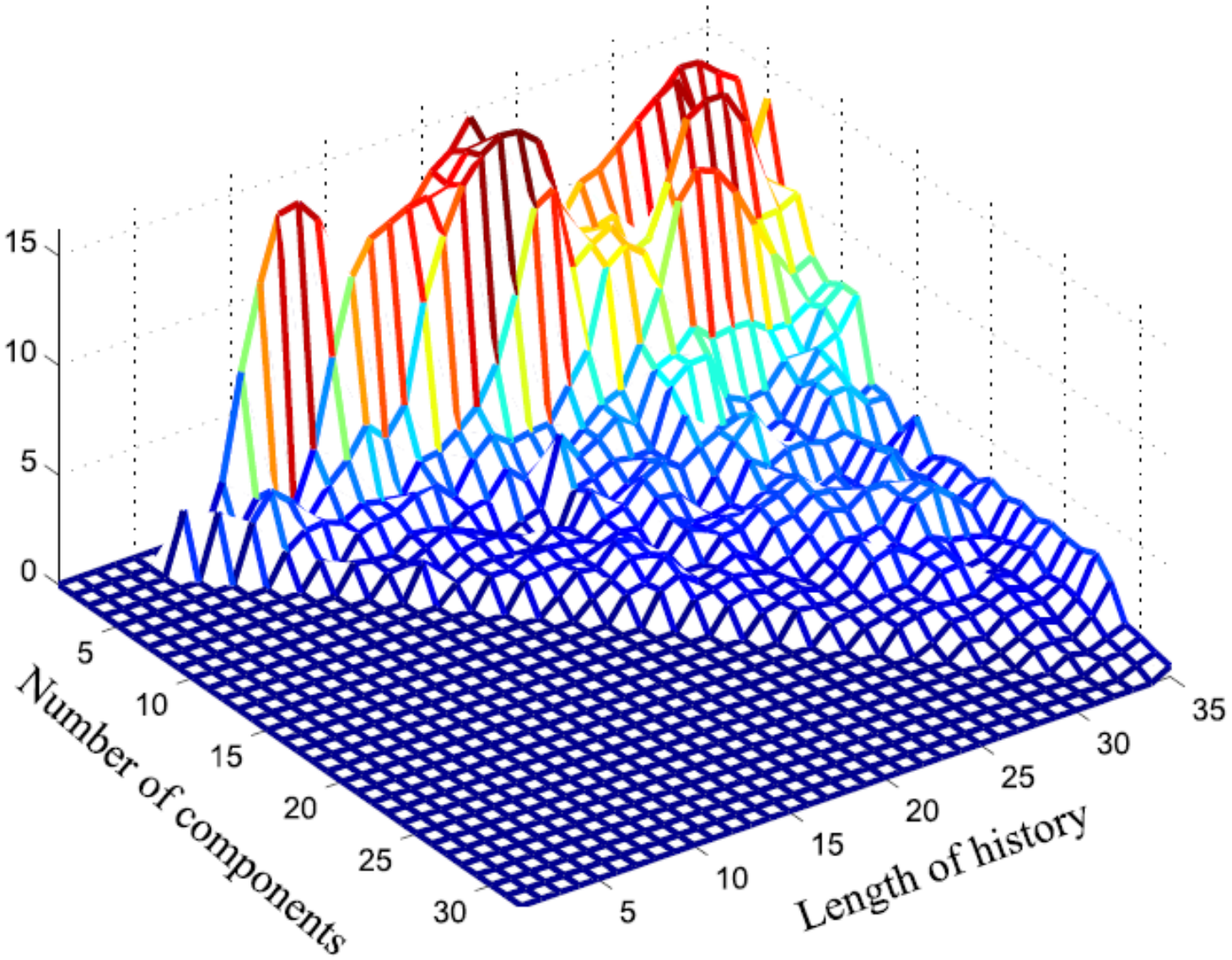


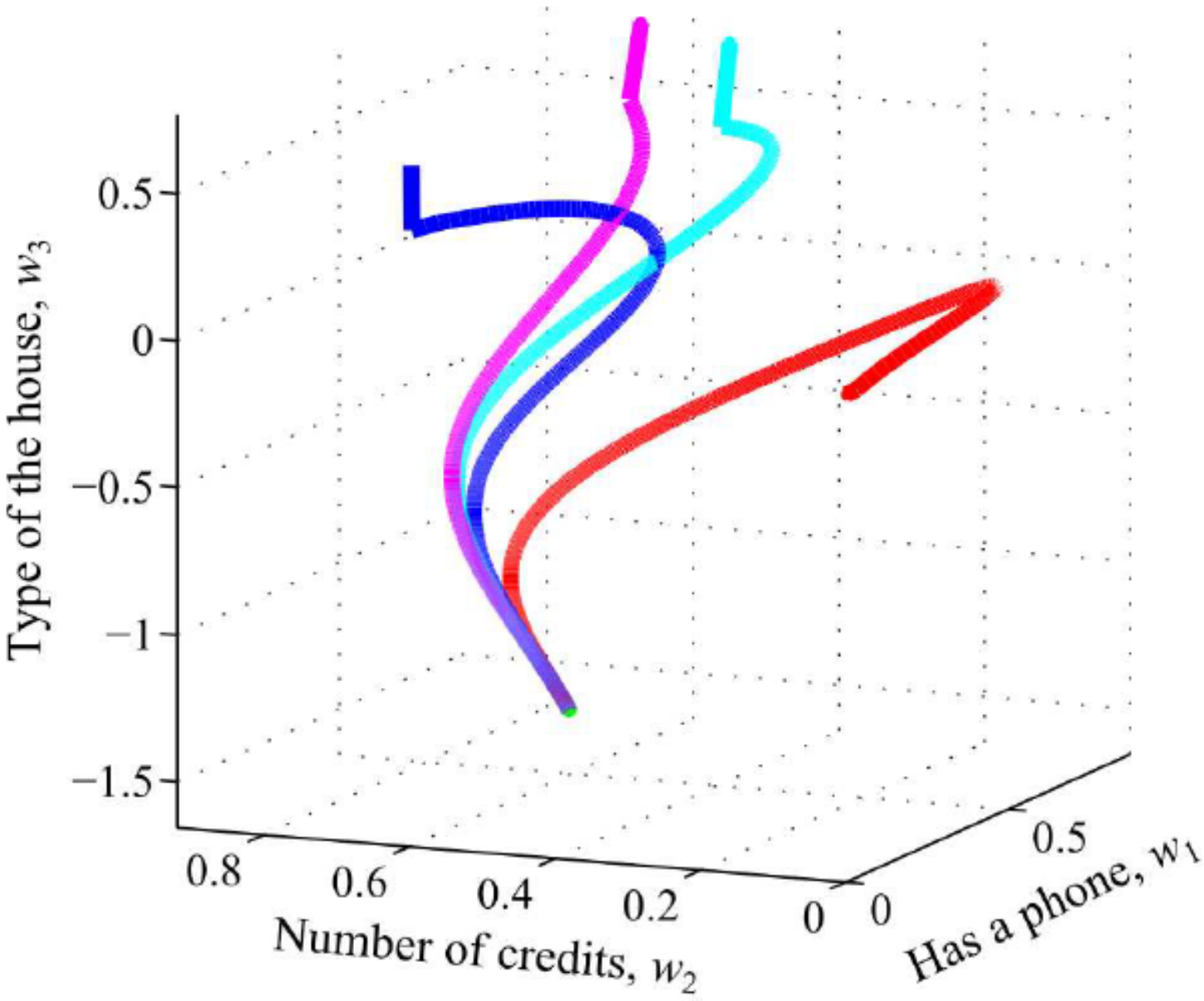

Пн
Вс
9.05
23.05
6.06
20.06
4.07
18.07
1.08
15.08
29.08
12.09
26.09

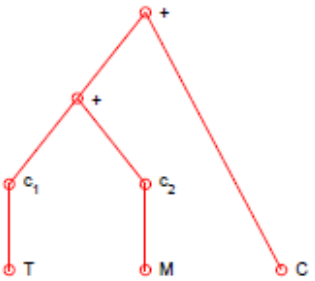

U
400
350
300
250
200
150
100
50
0
-50
-100
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
t, сек.
R_n
T_n
R_{n+1}
α_n


SSE
3
2
1
0
-1
30
25
20
15
10
5
Number components
35
30
25
20
15
10
5
Length

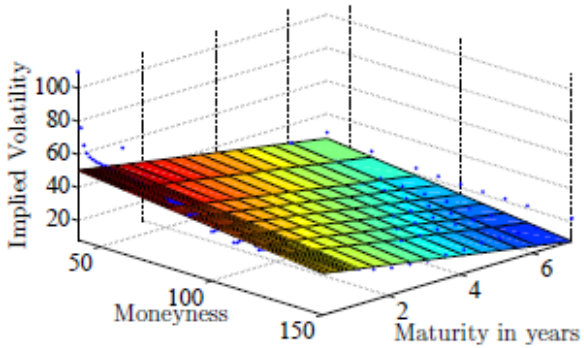

35
30
25
20
15
10
5
Length
5
10
15
20
25
30
Number components

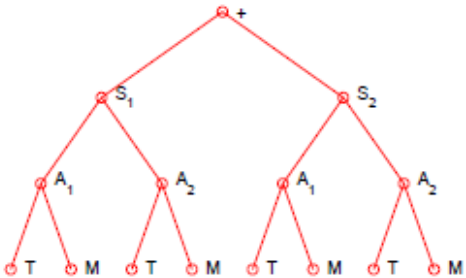

15
10
5
0
5
10
15
20
25
30
Number of components
5
10
15
20
25
30
35
Length of history


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

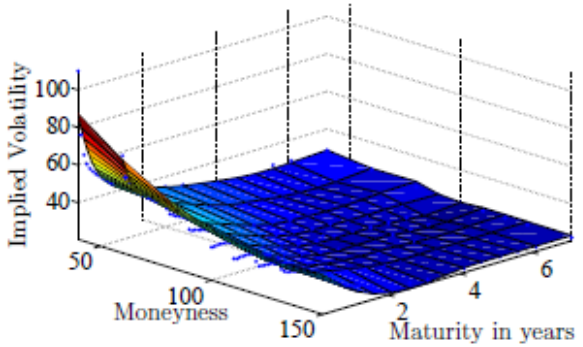

+
+
c1
c2
T
M
C

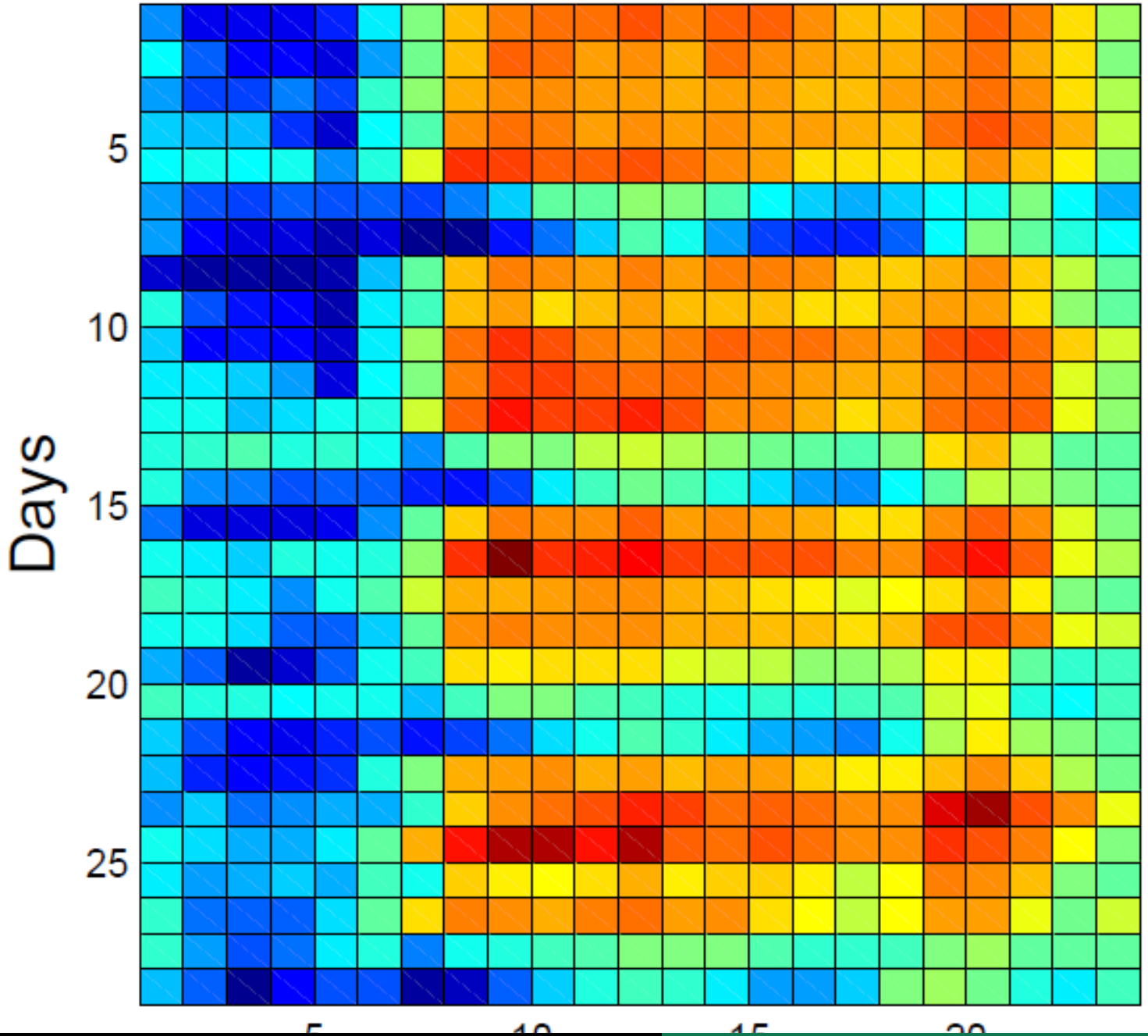
а)

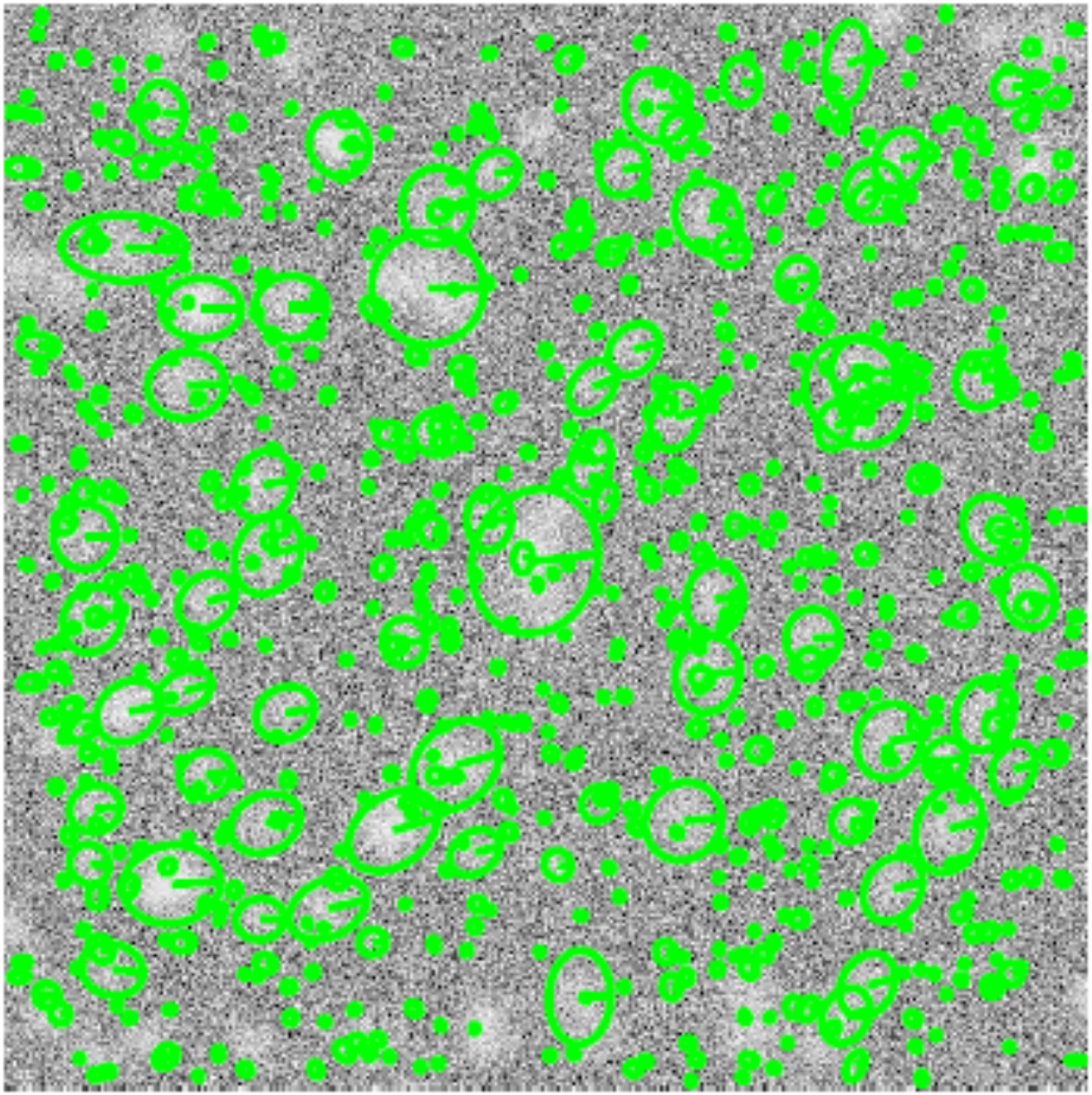

Implied Volatility
100
80
60
40
20
50
100
150
Moneyness
2
4
6
Maturity in years


+
S1
S2
A1
A2
A1
A2
T
M
T
M
T
M
T
M

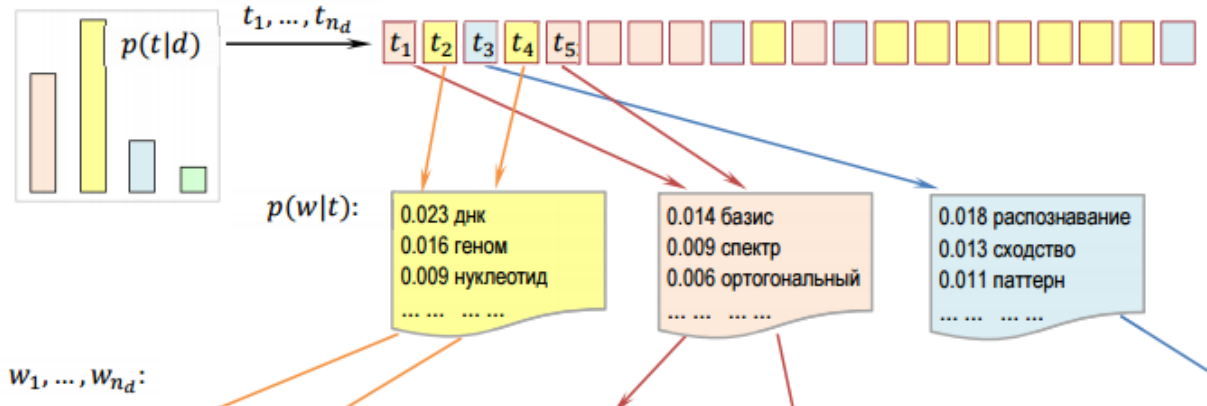
б)


Implied Volatility
100
80
60
40
50
100
150
Moneyness
2
4
6
Maturity in years


Days
5
10
15
20
25

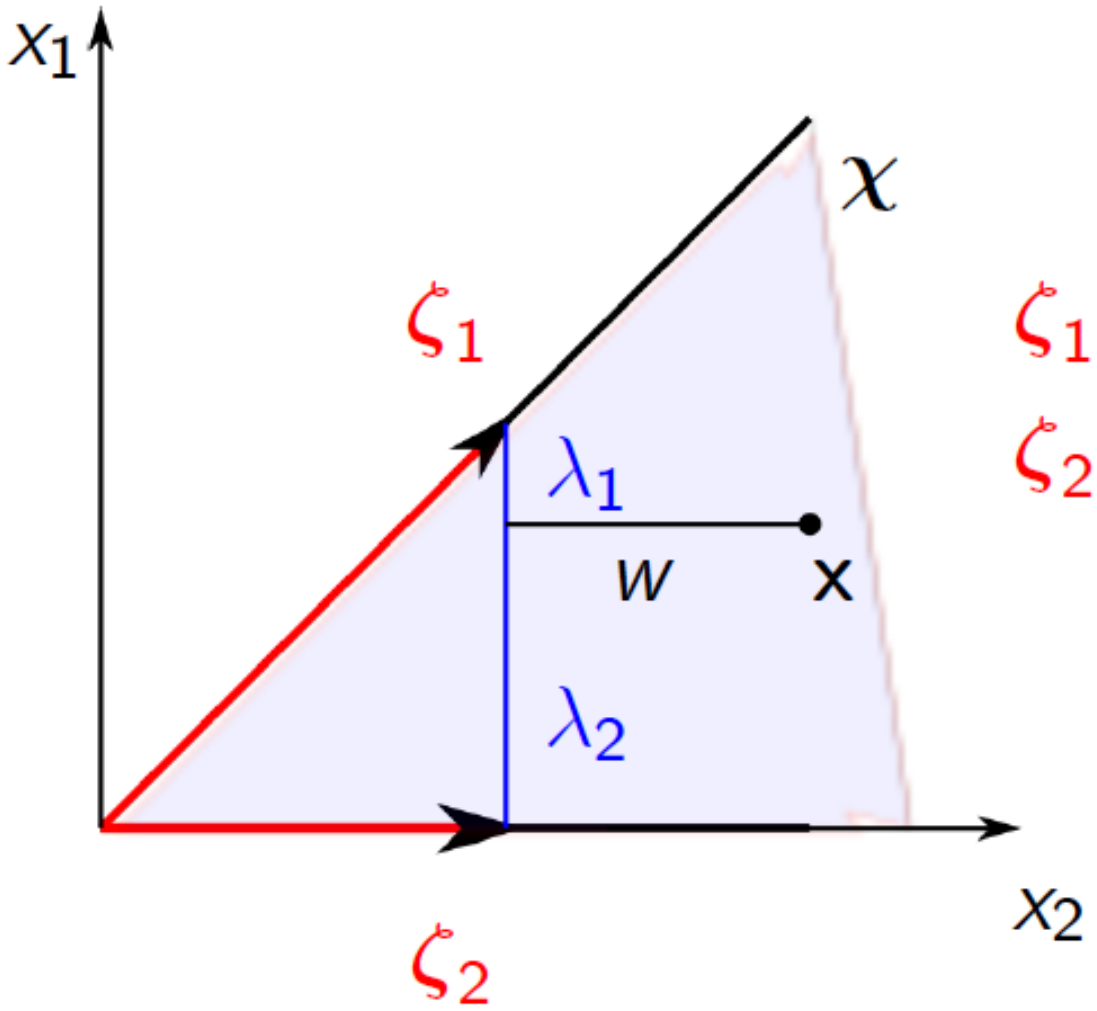


Тематическое моделирование и матричное разложение

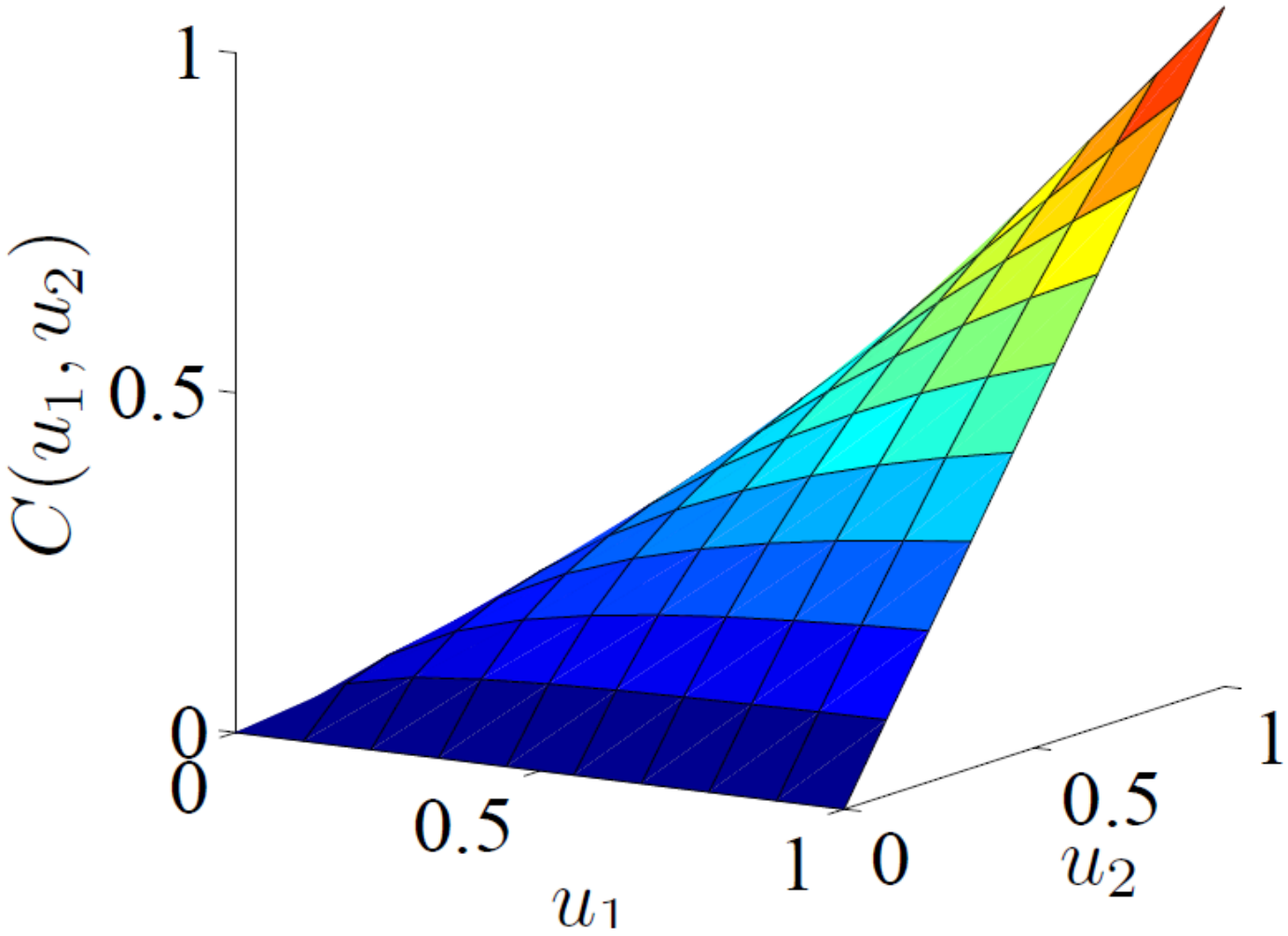


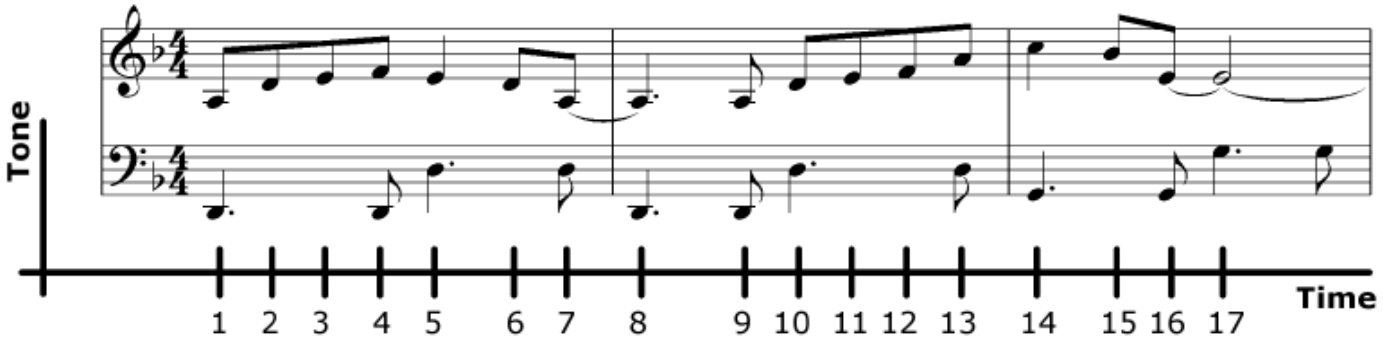
Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным 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).

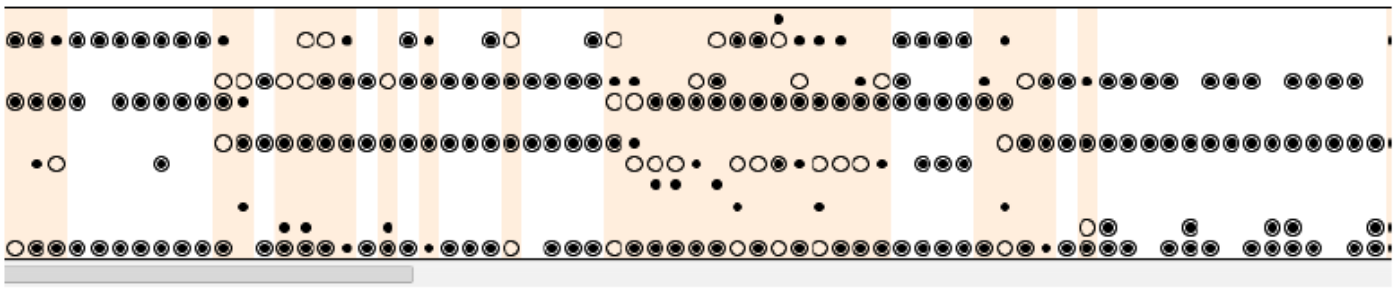
Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization // International Symposium On Learning And Data Sciences 2015.



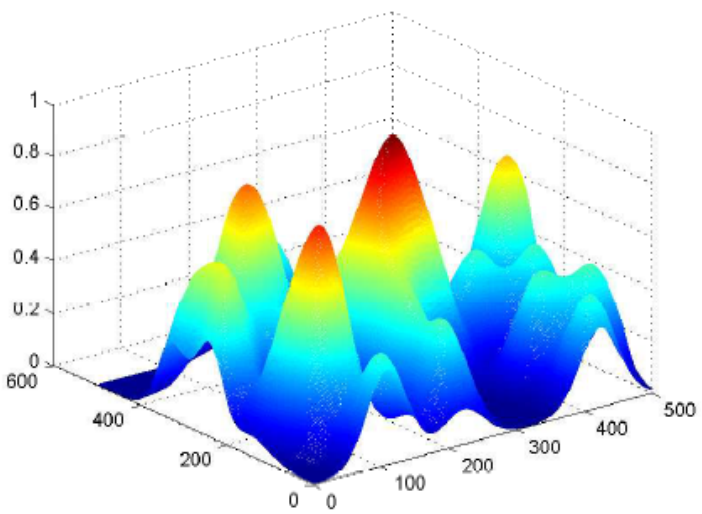
$$\zeta_1 = [1; 1],$$
$$\zeta_2 = [1; 0]$$


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

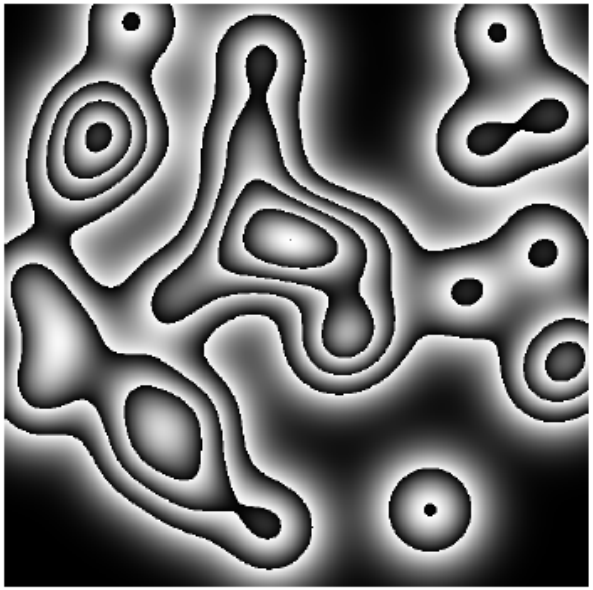




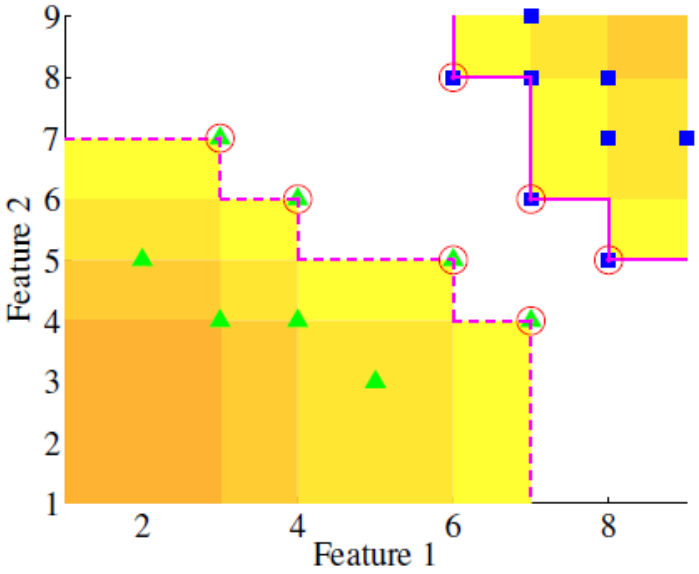
Пустые круги — истинные полутона, точки — предсказанные, ошибки подсвечены, горизонтальная ось — время.

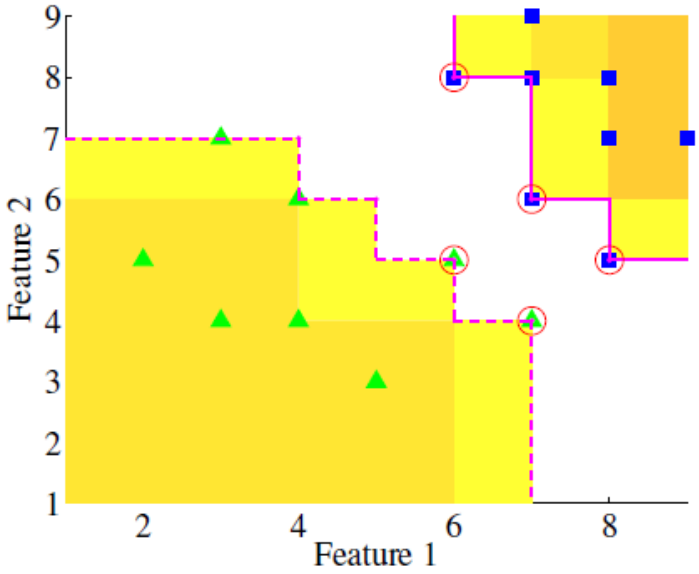


(а) Высота



(б) Фазовая составляющая без учета шума


9
8
7
6
5
4
3
2
1
Feature 2
2
4
6
8
Feature 1


9
8
7
6
5
4
3
2
1
Feature 2
2
4
6
8
Feature 1