Линейная регрессия. Дизайн экспериментов

Центр биоэлектрических интерфейсов, 19 декабря 2018 г.

Денис Деркач, Влад Белавин

Оглавление

Регрессия

Регрессия

Регрессия — метод изучения зависимости между откликом Y и регрессором X (признак, независимая переменная).

Один из способов оценить зависимость:

$$r(x) = \mathrm{E}\left(Y|X = x\right) = \int y\, f(y|x)dy.$$

Задача состоит в том, чтобы построить оценку $\hat{r}(x)$ функции $r(x)$ по данным

$$(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{X,Y},$$

где $F_{X,Y}$ — совместное распределение X и Y.

Стандартная линейная регрессия

Линейная регрессия

Линейная регрессия:

$$r(x) = \beta_0 + \beta_1 x.$$

Определение: простая линейная регрессия

Пусть $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, где ε_i — шум с мат. ожиданием $\mathbb{E}(\varepsilon_i | X_i) = 0$ и дисперсией $\mathbb{V}ar(\varepsilon_i | X_i) = \sigma^2$.

Оценивание параметров:

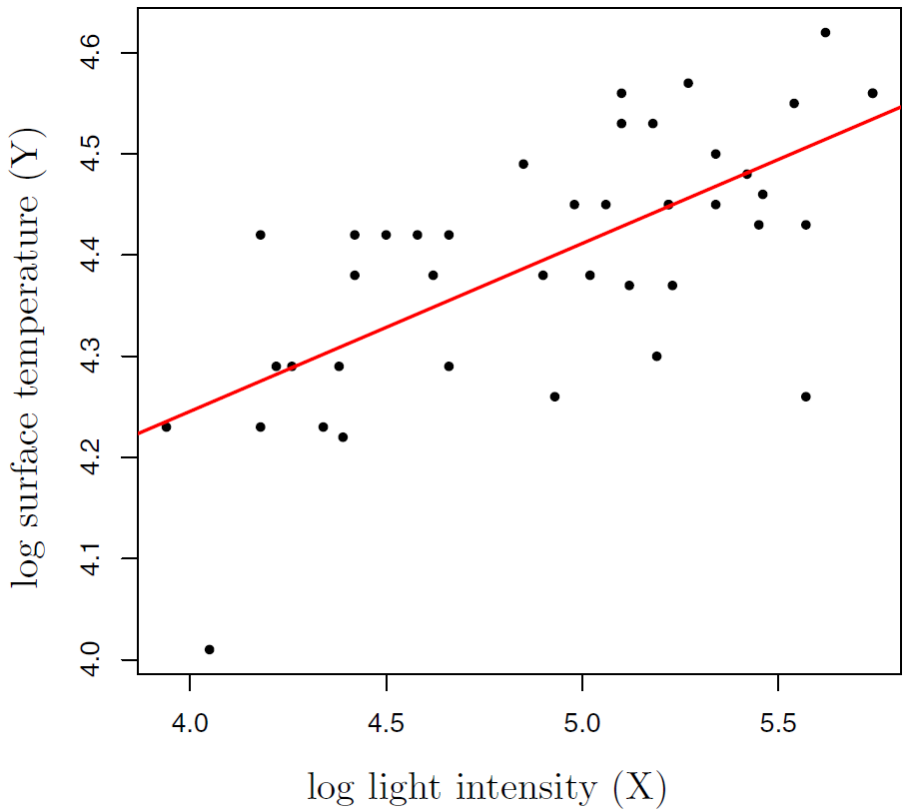
$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Предсказанные значения:

$$\hat{Y}_i = \hat{r}(X_i).$$

Примеры: линейная регрессия

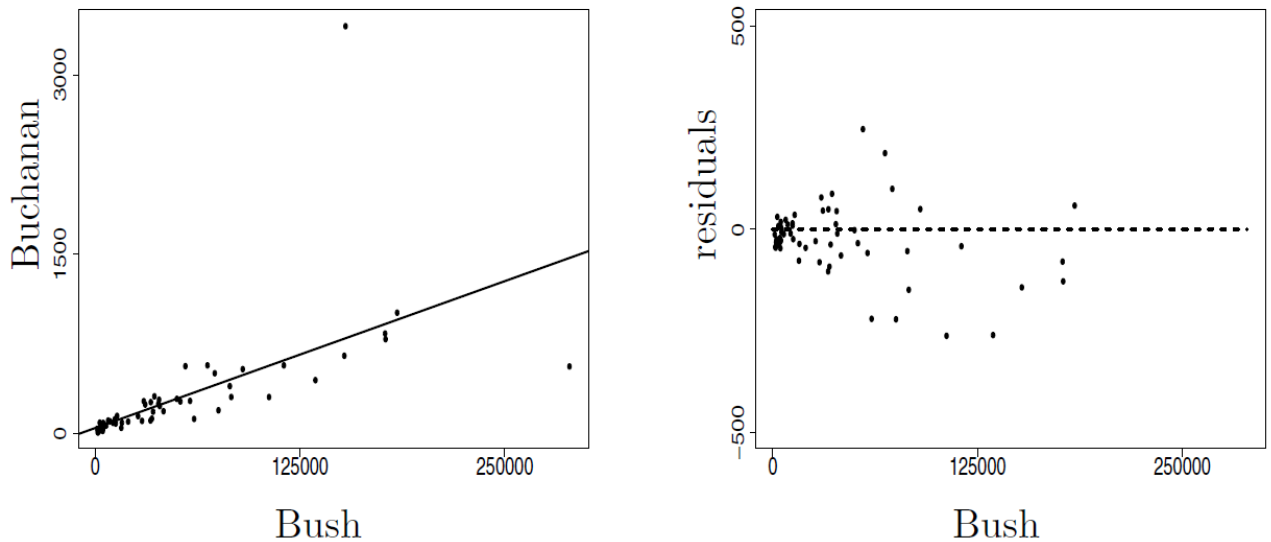
Данные о близлежащих звездах: оценка температуры звезды по её яркости.



Оценки равны: $\hat{\beta}_0 = 3.58$ и $\hat{\beta}_1 = 0.166 \Rightarrow \hat{r}(x) = 3.58 + 0.166x$.

Примеры: стандартная линейная регрессия

Голоса за Buchanan (Y) vs. голоса за Bush (X) во Флориде. Справа на графике указана величина отклонения от прогноза. Гауссовское распределение отклонений будет скорее всего говорить о том, что прогноз выбран правильно.



$$\hat{\beta}_0 = 66.0991, \;\; \hat{se}(\hat{\beta}_0) = 17.2926,$$

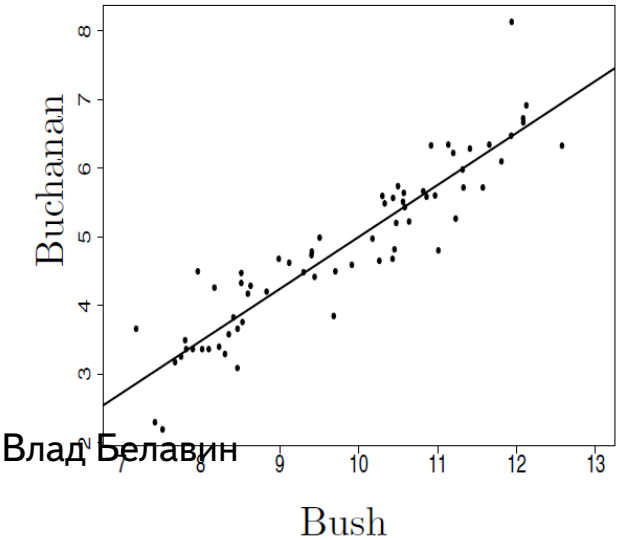
Примеры: стандартная линейная регрессия

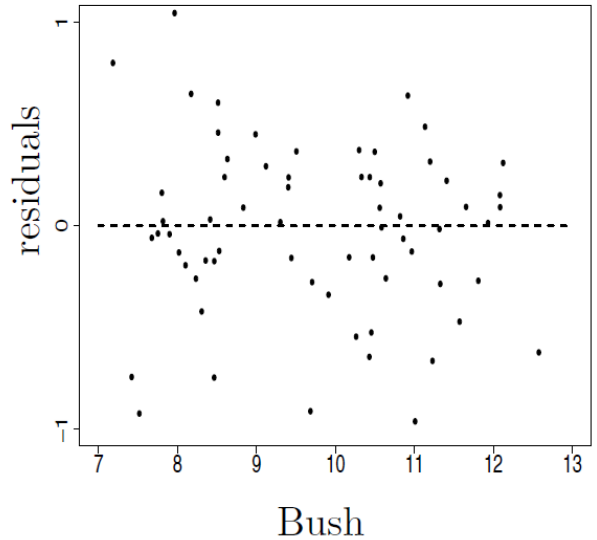
Если прологарифмировать данные, то остатки сильнее будут “напоминать” случайные числа:

$$\hat{\beta}_0 = -2.3298, \;\; \hat{se}(\hat{\beta}_0) = 0.3529,$$

$$\hat{\beta}_1 = 0.7303, \;\; \hat{se}(\hat{\beta}_0) = 0.0358,$$

$$\log(\text{Buchanan}) = -2.3298 + 0.7303 \log(\text{Bush}).$$





Метод наименьших квадратов

Остатки регрессии:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i).$$

Сумма квадратов остатков (RSS):

$$RSS = \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

Определение

$\hat{\beta}_0$ и $\hat{\beta}_1$ — оценки неизвестных параметров с помощью метода наименьших квадратов (МНК), если RSS для этих оценок минимальна.

Метод наименьших квадратов

Теорема

Оценки параметров β_0 и β_1 с помощью метода наименьших квадратов имеют вид

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2},$$

$$\hat{\beta}_0 = \overline{Y}_n - \hat{\beta}_1 \overline{X}_n.$$

При этом несмещенная оценка дисперсии шума σ^2 равна

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Свойства оценок МНК

Пусть $\hat{\beta}^T = (\hat{\beta}_0, \hat{\beta}_1)^T$ — оценка метода наименьших квадратов.
Тогда

$$\mathbb{E}(\hat{\beta}|X^n) = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix},$$

$$\mathbb{V}ar(\hat{\beta}|X^n) = \frac{\sigma^2}{ns_X^2}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n\\ -\overline{X}_n & 1\end{pmatrix}$$

при $s_X^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$

Таким образом,

$$\hat{se}(\hat{\beta}_0) = \frac{\hat{\sigma}}{s_X\sqrt{n}}\sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}, \qquad \hat{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{s_X\sqrt{n}}.$$

Свойства оценок МНК

Теорема

1. $\hat{\beta}_0 \xrightarrow{P} \beta_0,\ \hat{\beta}_1 \xrightarrow{P} \beta_1.$
2. $\frac{\hat{\beta}_0 - \beta_0}{\hat{se}(\hat{\beta}_0)} \rightsquigarrow \mathcal{N}(0,1),\ \frac{\hat{\beta}_1 - \beta_1}{\hat{se}(\hat{\beta}_1)} \rightsquigarrow \mathcal{N}(0,1).$
3. Приближенные доверительные интервалы размера $1-\alpha$ для параметров:
$$\hat{\beta}_0 \pm z_{\alpha/2}\hat{se}(\hat{\beta}_0) \text{ и } \hat{\beta}_1 \pm z_{\alpha/2}\hat{se}(\hat{\beta}_1).$$
4. Тест Вальда для проверки $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ имеет вид: H_0 отклоняется, если $|W| > z_{\alpha/2}$, где $W = \hat{\beta}_1 / \hat{se}(\hat{\beta}_1)$.

Пример: критерий Вальда

Замечание

Критерий Вальда для проверки $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ имеет вид $W = \frac{\hat{\beta} - \beta_0}{\hat{se}(\hat{\beta})}$.

Пример

(Выборы) Для регрессии (в логарифмическом масштабе) 95% доверительный интервал имеет вид

$$0.7303 + 2 \times 0.0358 = (0.66, 0.80).$$

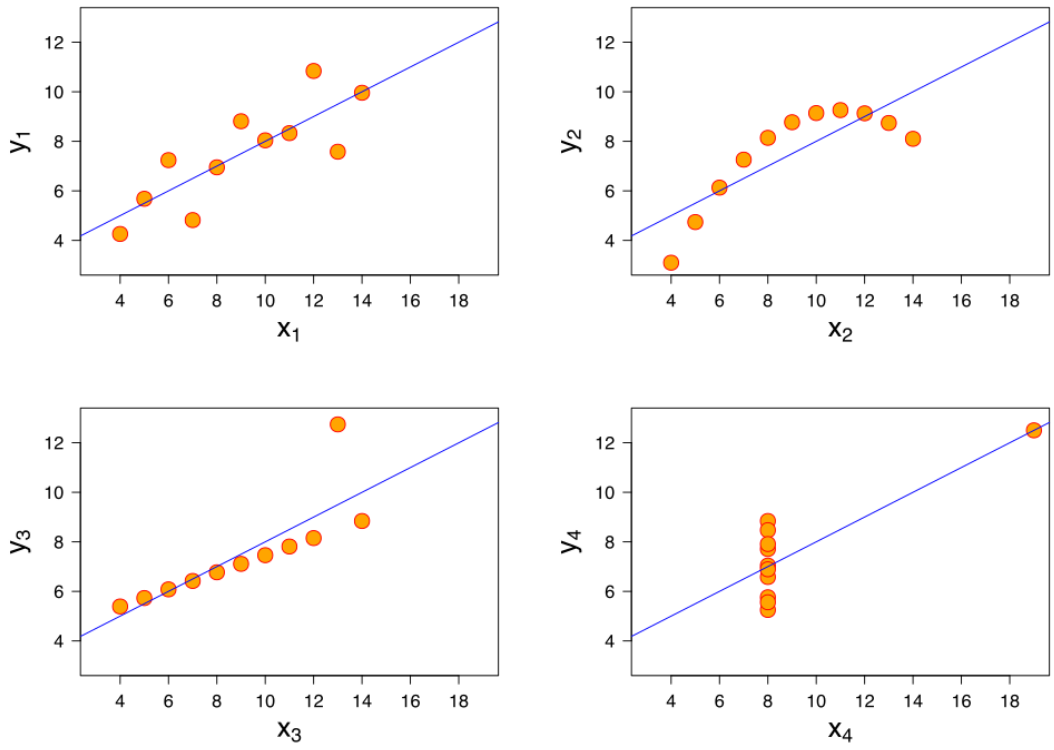
Статистика Вальда для проверки $H_0 : \beta_1 = 0$ против альтернативы $H_1 : \beta_1 \neq 0$ равна $|W| = |.7303 - 0|/.0358 = 20.40$. Причем $p - value$ равно $\mathrm{P}(|Z| > 20.40) \approx 0 \Rightarrow$ зависимость действительно существует.

Способы оценки качества регрессии

- Mean square (L2) loss (MSE): $\mathrm{MSE}(h, X^\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - h(x_i))^2$
- Root MSE: $\mathrm{RMSE}(h, X^\ell) = \sqrt{\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - h(x_i))^2}$
- Coefficient of determination: $R^2(h, X^\ell) = 1 - \frac{\sum_{i=1}^{\ell}(y_i - h(x_i))^2}{\sum_{i=1}^{\ell}(y_i - \mu_y)^2}$

  with $\mu_y = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i$
- Mean absolute error: $\mathrm{MAE}(h, X^\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell}|y_i - h(x_i)|$

Квартет Энскомба



Все 4 семейства имеют одинаковое средние, дисперсии, уравнения регрессии, R^2.

Datasaurus:https://bit.ly/2wtDgyFl

Множественная регрессия

Множественная регрессия

В этом случае данные имеют вид

$$(X_1, Y_1), \ldots, (X_i, Y_i), \ldots, (X_n, Y_n),$$

$$X_i = (X_{i1}, \ldots, X_{ik}) \in \mathbb{R}^k.$$

Модель имеет вид $(i = 1, \ldots, n)$

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i,$$

$$\mathrm{E}\left(\varepsilon_i | X_{1i}, \ldots, X_{ki}\right) = 0.$$

Чтобы включить нулевой коэффициент, обычно полагают $X_{i1} = 1$ при $i = 1, \ldots, n$.

Множественная регрессия

Модель может быть выписана:

$$
\begin{aligned}
y_1 &= \beta_1 x_{11} + \ldots \beta_d x_{d1} + \varepsilon_1, \\
y_2 &= \beta_1 x_{12} + \ldots \beta_d x_{d2} + \varepsilon_2, \\
&\ldots \\
y_\ell &= \beta_1 x_{1\ell} + \ldots \beta_d x_{d\ell} + \varepsilon_\ell,
\end{aligned}
$$

или в матричной форме:

$$
\begin{bmatrix} y_1 \\ \vdots \\ y_\ell \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{d1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1\ell} & x_{2\ell} & \ldots & x_{d\ell} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_\ell \end{bmatrix} \quad \longleftrightarrow \quad \boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}
$$

Множественная регрессия

Теорема

Предположим, что матрица $X^T X$ размера $k \times k$ невырожденная, тогда

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$
$$\mathbb{V}ar(\hat{\beta}|X^n) = \sigma^2 (X^T X)^{-1},$$
$$\hat{\beta} \approx \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1}).$$

Оценка функции регрессии имеет вид

$$\hat{r}(x) = \sum_{j=1}^{k} \hat{\beta}_j x_j,$$

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} \hat{\varepsilon}_i^2,$$

Доверительные интервалы: множественная регрессия

Приближенный доверительный интервал размера $1 - \alpha$ для β_j равен

$$\hat{\beta}_j \pm z_{\alpha/2} \hat{se}(\hat{\beta}_j),$$

где $\hat{se}^2(\hat{\beta}_j)$ — j-ый диагональный элемент матрицы $\hat{\sigma}^2(X^T X)^{-1}$.

Пример: множественная регрессия

Пример

Данные о преступлениях по 47 штатам США в 1960г.
http://lib.stat.cmu.edu/DASL/Stories/USCrime.html

Регрессор	$\hat{\beta}_j$	$\hat{se}(\hat{\beta}_j)$	t-value	p-value
Нулевой коэффициент	-589.39	167.59	-3.51	0.001
Возраст	1.04	0.45	2.33	0.025
Южный штат(да/нет)	11.29	13.24	0.85	0.399
Образование	1.18	0.68	1.7	0.093
Расходы	0.96	0.25	3.86	0.000
Труд	0.11	0.15	0.69	0.493
Количество мужчин	0.30	0.22	1.36	0.181
Численность населения	0.09	0.14	0.65	0.518
Безработные (14-24)	-0.68	0.48	-1.4	0.165
Безработные (25-39)	2.15	0.95	2.26	0.030
Доход	-0.08	0.09	-0.91	0.367

Метод оценивания на основе максимизации правдоподобия

Предположим, что $\varepsilon_i | X_i \sim \mathcal{N}(0, \sigma^2)$.

$$Y_i | X_i \sim \mathcal{N}(\mu_i, \sigma^2), \text{ где } \mu_i = \beta_0 + \beta_1 X_i.$$

Правдоподобие имеет вид

$$\prod_{i=1}^{n} f(X_i, Y_i) = \prod_{i=1}^{n} f_X(X_i) f_{Y|X}(Y_i|X_i) =$$

$$= \prod_{i=1}^{n} f_X(X_i) \times \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i) = \mathcal{L}_1 \times \mathcal{L}_2,$$

$$\mathcal{L}_1 = \prod_{i=1}^{n} f_X(X_i), \qquad \mathcal{L}_2 = \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i)$$

Метод оценивания на основе максимизации правдоподобия

Функция $\mathcal{L}_1$ не содержит параметры β_0 и β_1.

Рассмотрим $\mathcal{L}_2$ — условную функцию правдоподобия:

$$\mathcal{L}_2 \equiv \mathcal{L}(\beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i) \propto \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (Y_i - \mu_i)^2 \right.$$

$$\ell(\beta_0, \beta_1, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2. \quad (1)$$

ОМП $(\beta_0, \beta_1) \Leftrightarrow$ максимизация (1)$\Leftrightarrow$ минимизация RSS,

$$RSS = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2.$$

Метод оценивания на основе максимизации правдоподобия

Теорема

В предположении нормальности ОМП оценка совпадает с оценкой метода наименьших квадратов.

Максимизируя $\ell(\beta_0, \beta_1, \sigma)$ по σ, получаем ОМП оценку

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i \hat{\varepsilon}_i^2.$$

Прогнозирование

Регрессия

Стандартная линейная регрессия

Множественная регрессия

Прогнозирование

Выбор модели

Дизайн эксперимента

Заполнение пространства

- Random sampling
- Семплирование латинскими гиперкубами
- Полный факторный дизайн
- Halton sequence
- Свойства

Поверхность отклика

- Оптимальные Дизайны RSM
- Примеры
- RSM и категориальные переменные
- DoE для RSM

Адаптивный Дизайн

Прогнозирование

Модель — $\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, построенная по выборке данных $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Необходимо предсказать значение отклика Y_* при $X = x_*$:

$$\hat{Y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*.$$

$$\mathbb{V}ar(\hat{Y}_*) = \mathbb{V}ar(\hat{\beta}_0 + \hat{\beta}_1 x_*) = \mathbb{V}ar(\hat{\beta}_0) + x_*^2 \mathbb{V}ar(\hat{\beta}_1) + 2x_* Cov(\hat{\beta}_0, \hat{\beta}_1)$$

$\Rightarrow$ можно подсчитать $\hat{se}(\hat{Y}_*)$, используя в качестве оценки σ^2 величину $\hat{\sigma}^2$.

Прогнозирование

Теорема

Пусть

$$\hat{\xi}_n^2 = \hat{\sigma}^2 \left(\frac{\sum_{i=1}^n (X_i - X_*)^2}{n \sum_{i=1}^n (X_i - \overline{X})^2} + 1 \right).$$

Приблизительный prediction interval для Y_* размера $1 - \alpha$ имеет вид

$$\hat{Y}_* \pm z_{\alpha/2} \hat{\xi}_n.$$

Пример: прогнозирование

Пример

1. Выборы

$$\log(\text{Buchanan}) = -2.3298 + 0.7303 \log(\text{Bush}).$$

2. В Palm Beach за Bush отдали 152 954 голосов, а за Buchanan — 3 476.
3. В логарифмической шкале это составляет 11.93789 и 8.151045 соответственно.
4. Насколько вероятен этот исход в предположении, что модель верна?
   - Предсказание для Buchanan равно -2.3298 + 0.7303 * 11.93789 = 6.388441.
5. Существенно ли это меньше, чем мы наблюдаем на практике?
   - $\hat{\xi} = 0.093775$ и 95% доверительный интервал имеет вид

Выбор модели

Выбор модели

Бритва Оккама — не надо “плодить” сущности. Много переменных приводят к большой дисперсии прогноза, но маленькому смещению, и наоборот.

При выборе подходящей модели возникают две задачи:

1. выбор целевой функции для характеризации качества используемой модели;
2. поиск оптимальной модели согласно выбранному критерию качества.

Обозначения

Пусть $S \subset \{1, \ldots, k\}$ — подмножество регрессоров. Тогда

- β_S — коэффициенты при соответствующих регрессорах, $\hat{\beta}_S$ — их оценки;
- X_S — подматрица матрицы типа X в соответствии с данным подмножеством регрессоров;
- $\hat{r}_S(x)$ — оцененная функция регрессии, $\hat{Y}_i(S) = \hat{r}_S(X_i)$ — предсказанные значения.

Риск прогноза

Риск прогноза:

$$R(S) = \sum_{i=1}^{n} \mathrm{E}\,(\hat{Y}_i(S) - Y_i^*)^2,$$

где $Y_i^* = X_i\beta$ — истинное значения выхода для X_i.

Задача состоит в выборе подмножества S, которое минимизирует $R(S)$.

Оценка риска прогноза

Оценка риска прогноза (ошибка на обучающей выборке):

$$\hat{R}_{tr}(S) = \sum_{i=1}^{n} (\hat{Y}_i(S) - Y_i)^2.$$

Теорема

Оценка риска прогноза смещена по сравнению с реальным значением риска прогноза:

$$bias\big(\hat{R}_{tr}(S)\big) = \mathbb{E}\hat{R}_{tr}(S) - R(S) = \sum_{i=1}^{n} \Big(\sigma_i^2 - 2\operatorname{cov}\left(\hat{Y}_i(S), Y_i\right)\Big).$$

Оценка риска прогноза

- Причина в том, что данные использовались дважды — для оценки параметров и для оценки риска прогноза.
- Если параметров много, то $\operatorname{cov}\left(\hat{Y}_i(S), Y_i\right)$ принимает большое значение.
- При этом прогноз на данных, отличных от данных в обучающей выборке, может оказаться существенно хуже!

Статистика C_p Mallow

Статистика C_p Mallow:

$$\hat{R}(S) = \hat{R}_{tr}(S) + 2|S|\hat{\sigma}^2,$$

где $|S|$ — число регрессоров, $\hat{\sigma}^2$ — оценка дисперсии шума σ^2, полученная по полной модели (т.е. с включением всех регрессоров).

Критерий включает оценку риска прогноза на обучающей выборке и “сложность” модели (регуляризация).

AIC

AIC (Akaike information criterion):

$$AIC(S) = \ell_S - |S| \to \max_S,$$

где $\ell_S = \ell_S(\hat{\beta})$ — логарифм правдоподобия модели, где в качестве неизвестных параметров были подставлены их оценки, полученные с помощью максимизации $\ell_S(\beta)$.

В линейной регрессии в случае нормальных ошибок (шум берется равным оценке, полученной по полной модели) максимизация AIC эквивалента минимизации C_p.

Кросс-проверка

Оценка риска с помощью кросс-проверки (cross-validation; leave-one-out):

$$\hat{R}_{CV}(S) = \sum_{i=1}^{n} (\hat{Y}_{(i)} - Y_i)^2,$$

где $\hat{Y}_{(i)}$ — предсказание значения Y_i, полученное по модели, параметры которой оценены на обучающей выборке без i входа.

$$\hat{R}_{CV}(S) = \sum_{i=1}^{n} \frac{(\hat{Y}_i - Y_i)^2}{1 - U_{ii}(S)},$$

$$U(S) = X_S (X_S^T X_S)^{-1} X_S^T.$$

K-кратная кросс-проверка

1. Данные случайным образом делятся на k непересекающихся подвыборок (часто берут $k = 10$).
2. По одной подвыборке за раз удаляется (с возвращением), по остальным происходит оценка параметров.
3. Риск полагается равным $\sum_i (\hat{Y}_i - Y_i)^2$ (сумма берется по наблюдениям из удаленной подвыборки, данные оцениваются с помощью полученной модели).
4. Процесс повторяется для остальных подвыборок, после чего полученная оценка риска усредняется.

Для линейной регрессии оценка на основе коэффициента C_p Mallow и оценка на основе K-кратной кросс-проверки зачастую совпадают. В более сложных случаях кросс-проверка работает лучше.

BIC

BIC (Bayesian information criterion):

$$BIC(S) = \ell_S - \frac{|S|}{2}\log n \to \max_S .$$

Этот функционал имеет байесовскую интерпретацию.

- Пусть $\mathcal{S} = \{S_1, \ldots, S_m\}$ — множество возможных моделей.
- Допустим, что априорное распределение имеет вид $\mathrm{P}(S_j) = 1/m$.
- Также предположим, что параметры внутри каждой модели имеют некоторое “гладкое” априорное распределение.
- Можно показать, что апостериорная вероятность модели примерно равна

$$\mathrm{P}(S_j|\text{выборка}) \approx \frac{\exp\big(BIC(S_j)\big)}{\sum_{r=1}^{m}\exp\big(BIC(S_r)\big)}.$$

BIC

Таким образом, выбор модели с наибольшим BIC эквивалентен выбору модели с наибольшей апостериорной вероятностью.

BIC также можно интерпретировать с точки зрения теории минимальной длины описания информации: BIC обычно “выбирает” модели с меньшим числом параметров.

Перебор моделей

- Если в модели максимальное количество регрессоров равно k, то существует 2^k всевозможных моделей.
- В идеале необходимо “просмотреть” все модели, для каждой найти значение критерия качества и выбрать наилучшую согласно этому критерию.
- При большом количестве регрессоров для уменьшения трудоемкости используют регрессию методом включений, исключений или включений-исключений.

Метод включений / метод исключений

- Включения:
  - на первом шаге регрессоров нет вообще;
  - далее добавляется регрессор, для которого критерий качества максимальный и т.д.
- Исключения:
  - на первом шаге количество регрессоров максимальное;
  - на каждом шаге удаляется регрессор, исключение которого приводит к максимальному значению критерия качества.

Пример: метод исключений

Пример

Данные о преступлениях. Используем критерий AIC, что эквивалентно минимизации C_p Mallow.

В модели с полным набором регрессоров AIC = -310.37. В порядке убывания AIC при удалении каждой из переменных равен:

Численность населения (AIC = -308), Труд (AIC = -309), Южный штат (AIC = -309), Доход (AIC = -309), Количество мужчин (AIC = -310), Безработные I (AIC = -310), Образование (AIC = -312), Безработные II (AIC = -314), Возраст (AIC = -315), Расходы (AIC = -324).

Таким образом, имеет смысл удалить переменную “Население”.

Пример: метод исключений

Южный штат (AIC = -308), Труд (AIC = -308), Доход (AIC = -308), Количество мужчин (AIC = -309), Безработные I (AIC = -39), Образование (AIC = -310), Безработные II (AIC = -313), Возраст (AIC = -313), Расходы (AIC = -329).
Удаляем переменные до тех пор, пока не удастся больше получить увеличения AIC.

Уровень преступности = 1.2 Возраст + 0.75 Образование + 0.87 Расходы + 0.34 Количество мужчин − 0.86 Безработные I + 2.31 Безработные II.

Замечание: не дан ответ на то, какие переменные вызывают рост уровня преступности!

Дизайн эксперимента

Мотивация

Типы задач:

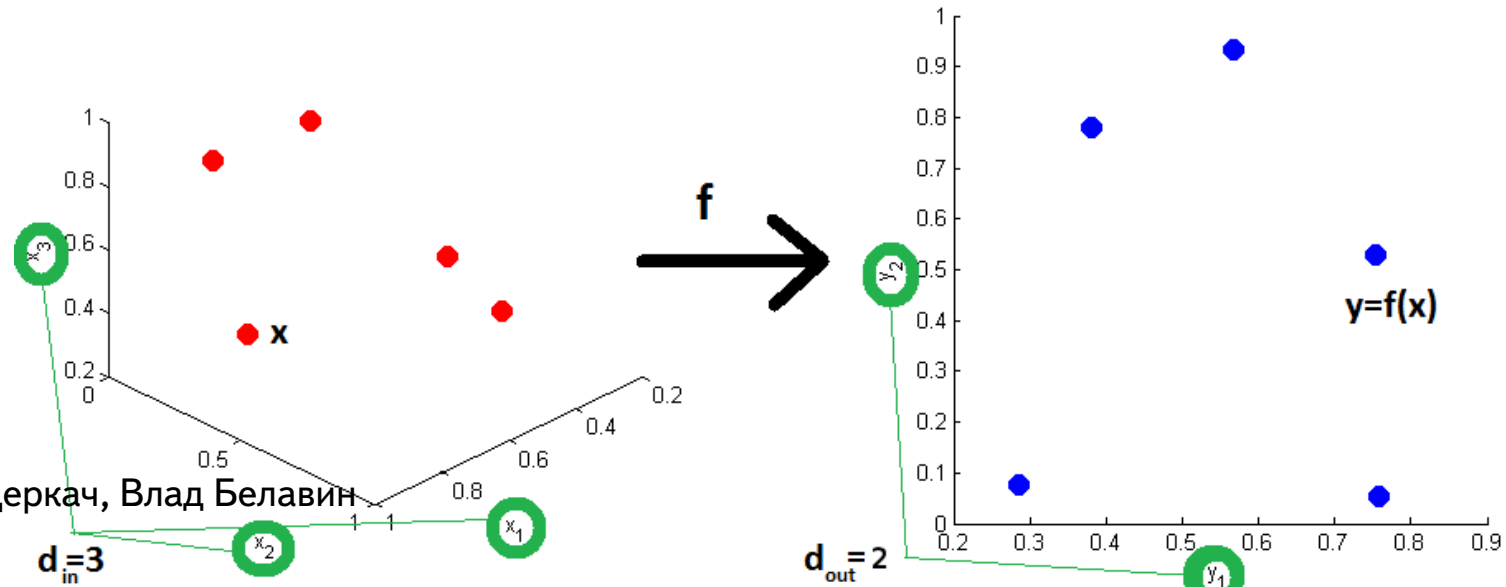
- сравнительный эксперимент - сравнение двух разных моделей;
- наблюдательный эксперимент - упрощение модели;
- построение поверхности отклика;
- регрессия.

Во всех случаях стоимость вычислений может быть достаточно высокой

Данные

Определение

- $\mathbb{X}$ — d_{in}-мерное пространство.
- $\mathbf{x} \in \mathbb{X}$ — d_{in}-мерный вектор, описывающий дизайн объекта.
- $\mathbf{y}$ — d_{out}-мерный вектор характеристик объекта.
- Точная зависимость $f : \mathbf{x} \mapsto \mathbf{y}$.



Суррогатное моделирование

Определение

- $X = \{\mathbf{x}_i\}_{i=1}^N, \mathbf{x}_i \in X$ называется планом эксперимента.
- $D = (X, Y = f(X)) = \{(\mathbf{x}_1, \mathbf{y}_1 = f(\mathbf{x}_1)), \ldots, (\mathbf{x}_N, \mathbf{y}_N = f(\mathbf{x}_N))\}$ — обучающая выборка размера N.

Пример

	Вход $\leftrightarrow$ компоненты $\mathbf{x}$			Выход ($\mathbf{y}$)
	Angle of attack	...	Mach	Lift coeff.
Объект 1	0.5	$\ldots$	0.60	0.42
Объект 2	1.4	$\ddots$	0.77	0.62
$\vdots$	...	$\ddots$	$\ddots$	...
Объект N	3.2	$\ldots$	0.66	0.55

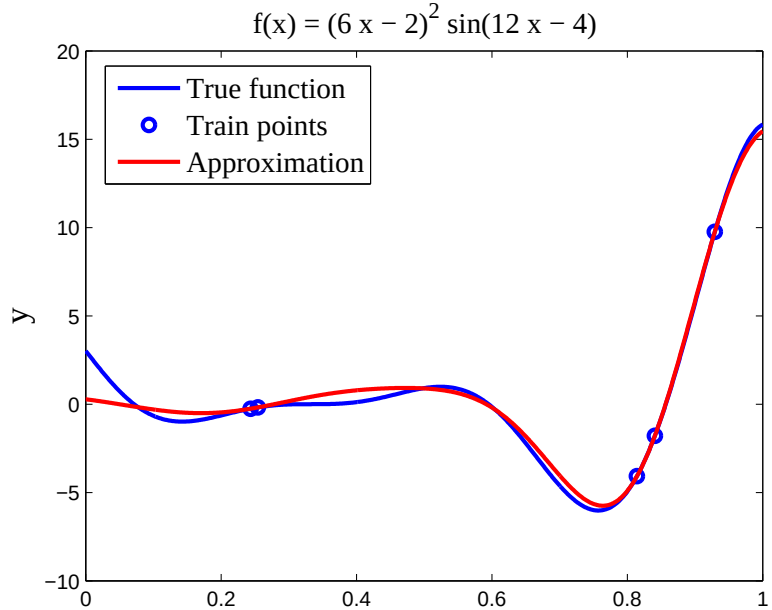
Эксперимент

Определение

$\hat{f}(\mathbf{x}) \approx f(\mathbf{x})$ суррогатная модель (функция), построенная по обучающей выборке $D = (X, Y = f(X))$.

Требование к $\hat{f}$:

- Легко и эффективно оценивается.
- “Близка” к f по значениям не только из D.



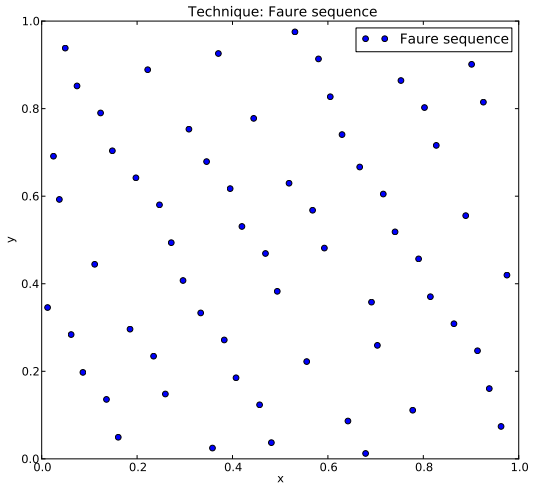
Заполнение пространства

Space-filling Design of Experiments

Определение

Space-filling DoE — это равномерное сэмплирование в гиперкубе.

Предположение: мы ничего не знаем о суррогатной модели и указанных зависимостях. Поэтому будем заполнять пространство дизайна равномерно.



Random sampling

Метод состоит в равномерной генерации точек в гиперкубе.

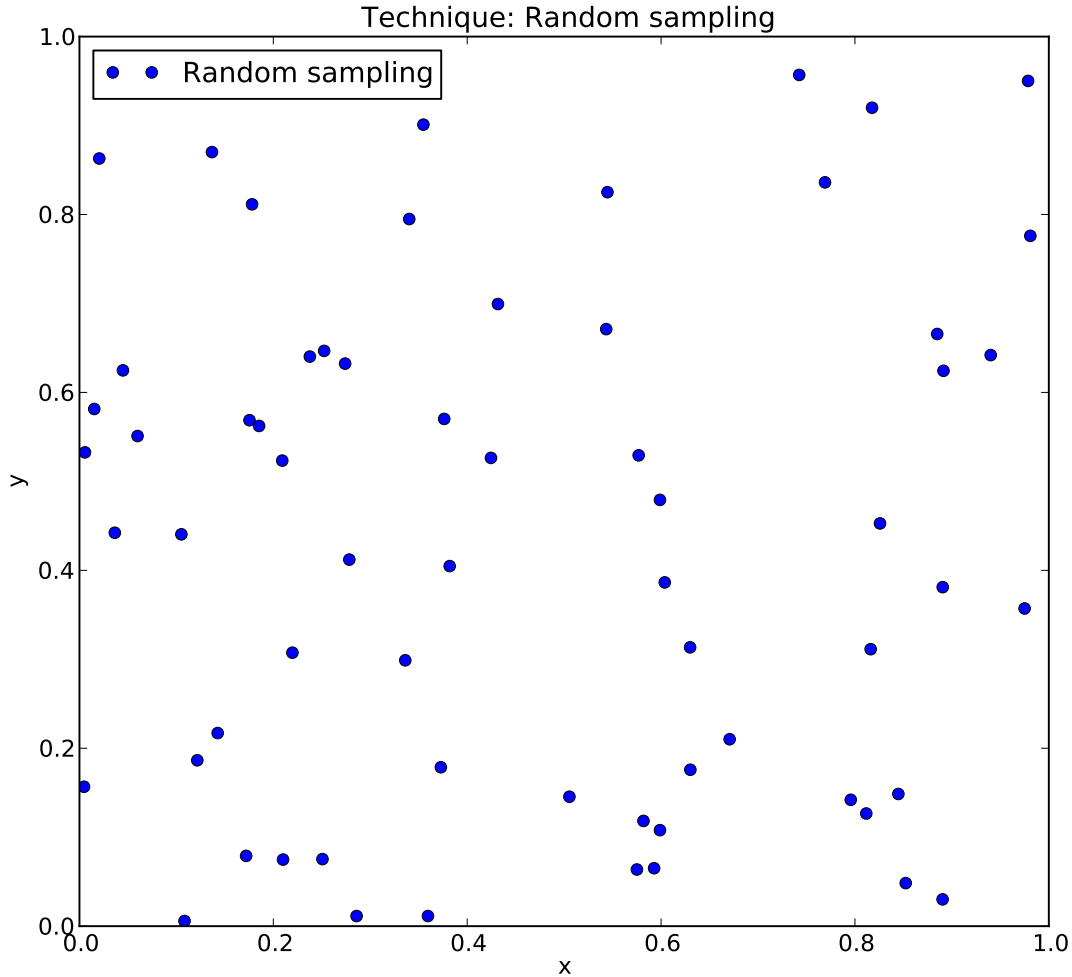
Преимущества

- Универсальность и гибкость.
- Всегда можно расширить добавлением точек.

Недостатки

- Нет гарантий равномерности заполнения.

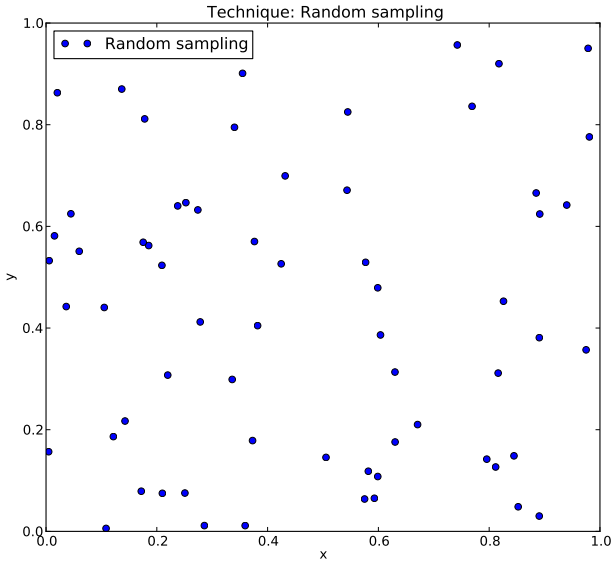
Random sampling

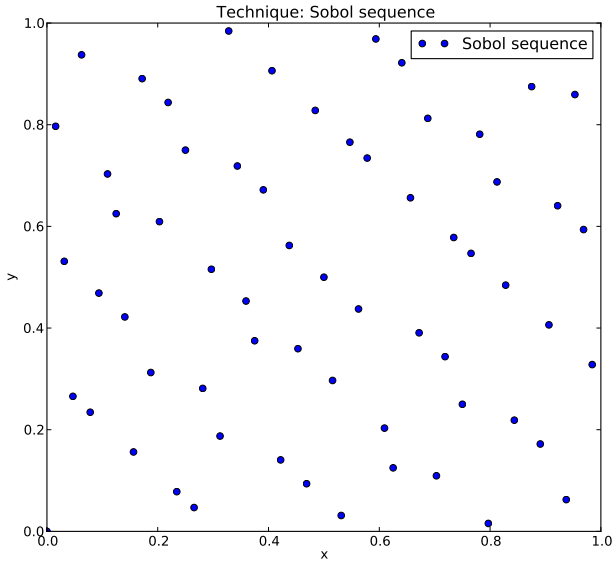


Random sequence ($N = 64$, $d_{in} = 2$, $\mathbb{X} = [0, 1] \times [0, 1]$)

Свойства: равномерность

Равномерность является важным свойством space-filling DoE.





Ожидаемые свойства:

- Отсутствие ”дырок” в пространстве дизайна.
- Равномерность малоразмерных проекций.

Критерии равномерности

Пусть $\mathbb{X} = [0, 1]^d$.

- Discrepancy

$$d(X) = \sup_{0 \le u_k < v_k < 1} \left| \frac{\#(X \cap P_{u,v})}{N} - |P_{u,v}| \right|, \quad P_{u,v} = \bigotimes_k [u_k, v_k),$$

- Минимаксное расстояние между точками

$$\rho(X) = \max_i \min_{j, j \neq i} \|\mathbf{x}_i - \mathbf{x}_j\|$$

- ϕ-метрика

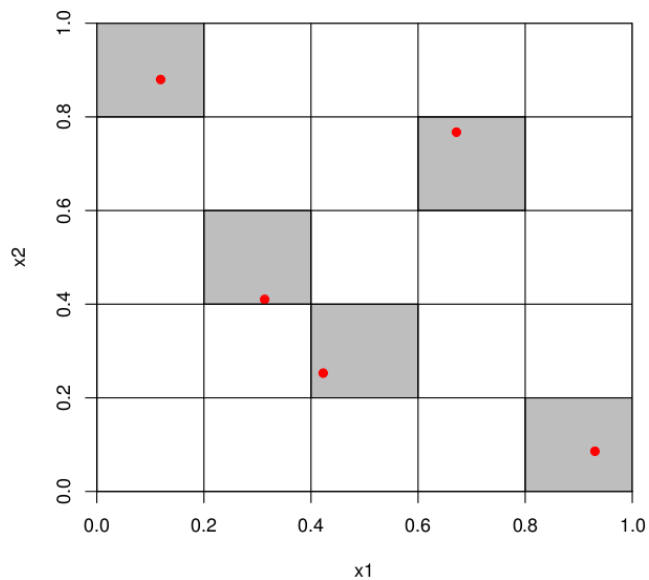
$$\phi_p(X) = \left(\sum_{i<j} \|\mathbf{x}_i - \mathbf{x}_j\|^{-p} \right)^{1/p}$$

Чем равномернее, тем ниже значение метрики.

Латинские гиперкубы

Определение

Семплирование латинским гиперкубом выполняется при помощи разделения значений каждой компоненты дизайна на N равных **интервалов**, в каждый из которых попадёт по одной точке.



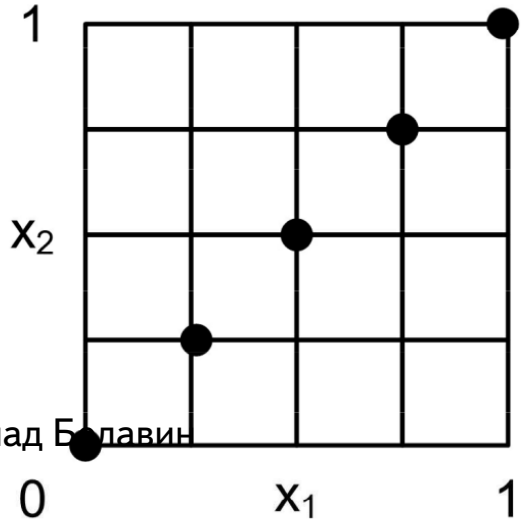
$$N = 5, d_{in} = 2, X = [0, 1] \times [0, 1]$$

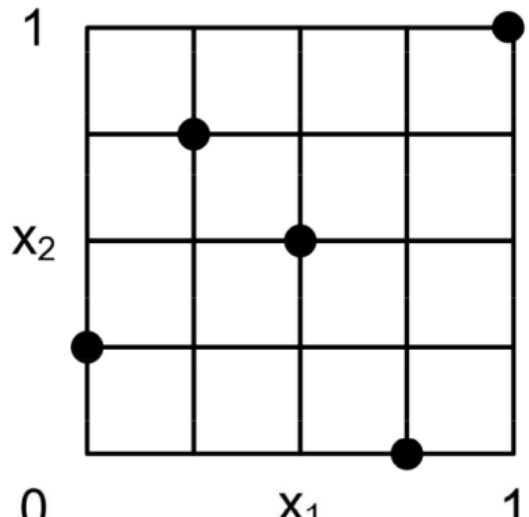
Оптимизированный LHS (OLHS)

Определение

Optimized Latin hypercube sampling это улучшенная версия LHS.

- LHS может иногда давать нежелательный результат.
- OLHS генерирует множество LHS дизайнов, а потом выбирает лучший по равномерной метрике.
- Это сильно увеличивает робастность, но требует много времени.





Optimized LHS (OLHS)

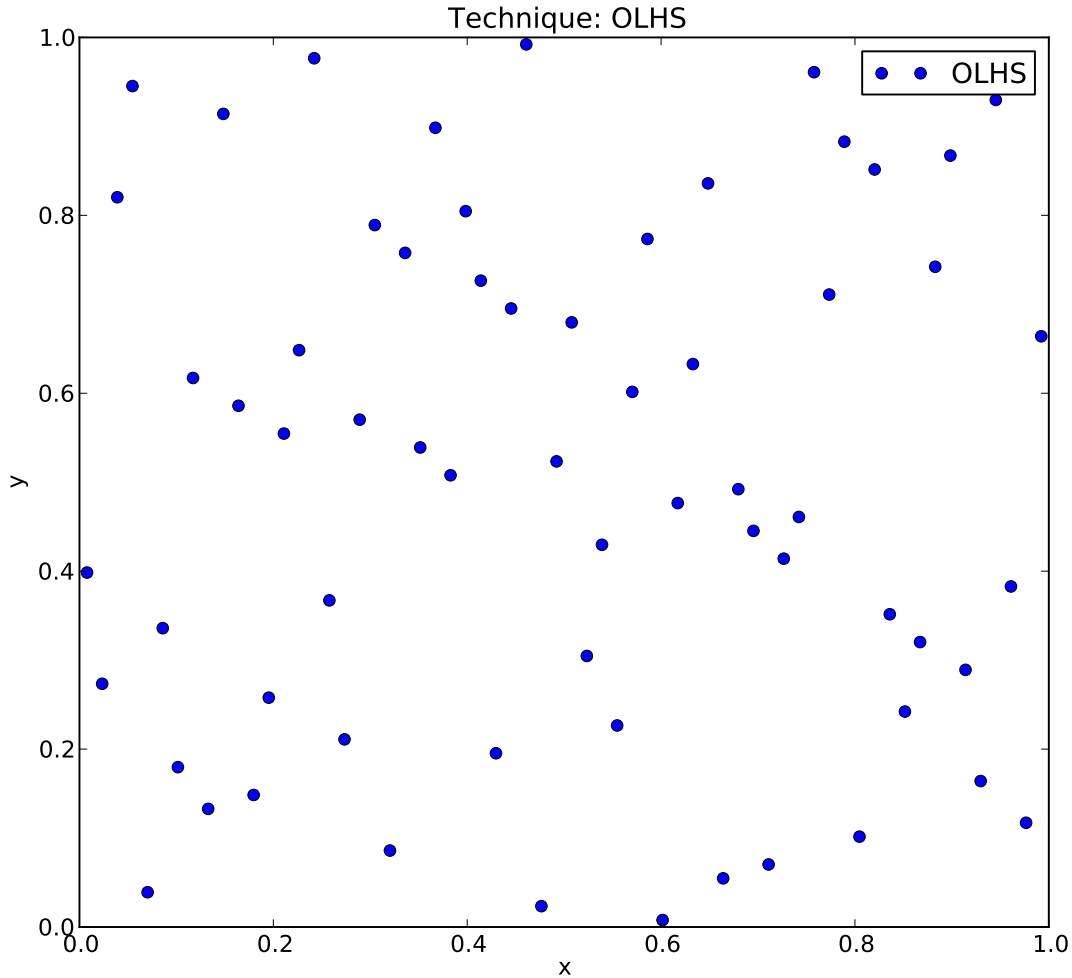
Преимущества

- Проекции на оси достаточно равномерны.
- Простота.

Недостатки

- Оптимизация очень медленная.
- Практически невозможно дополнить множество дизайна без нарушения свойств латинского гиперкуба (так как заранее заданное число N определяет число интервалов).

Optimized LHS (OLHS)



$N = 64, d_{in} = 2, X = [0, 1] \times [0, 1]$

Полный факторный план эксперимента

Определение

Полный факторный план эксперимента — все возможные комбинации уровней переменной дизайна, то есть все прямые произведения $X = \prod_{j=1}^{d_{in}} L_j = L_1 \times L_2 \times \ldots \times L_{d_{in}}$ конечных множеств L_1, $L_2, \ldots, L_{d_{in}}$, где L_i — это конечное множество из n_i точек интервала.

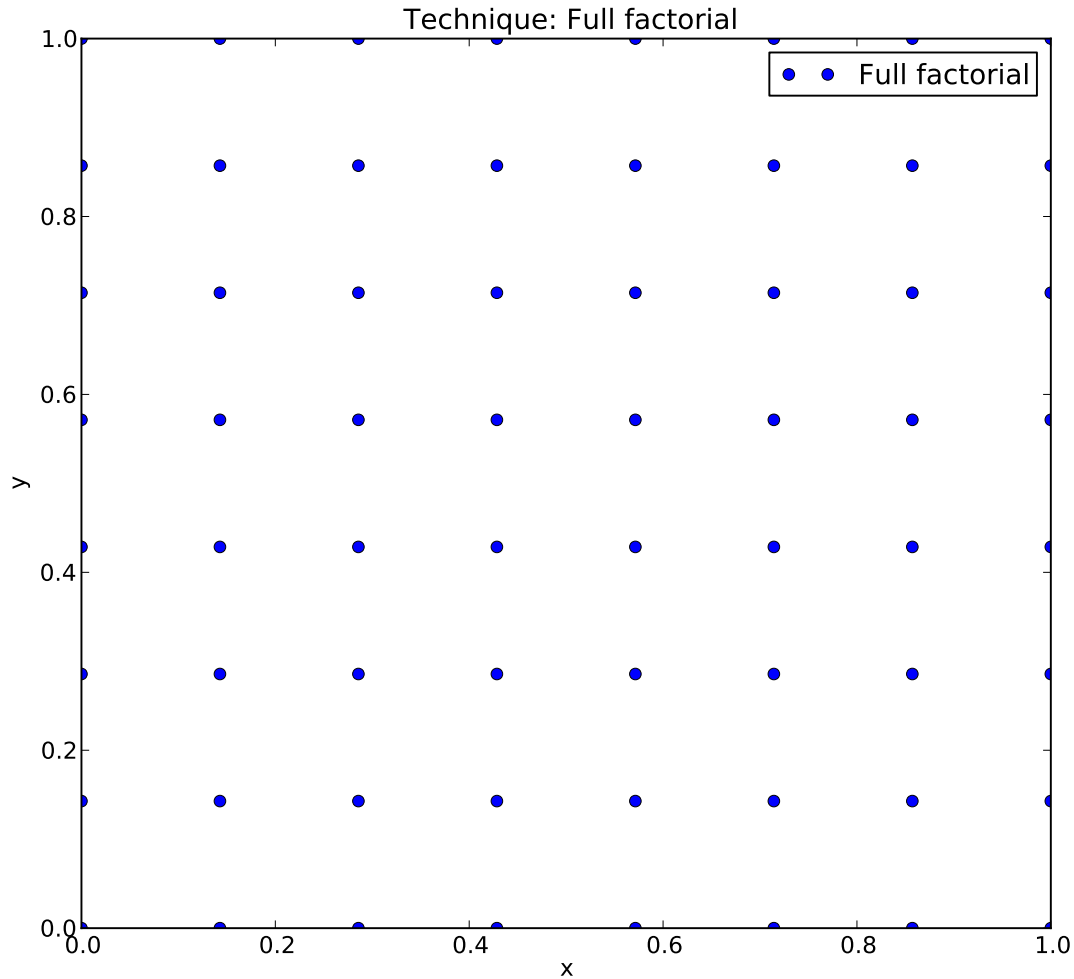
Преимущества

- Хорошо заполняет пространство.
- Очень быстро генерируется.

Недостатки

- Требуемое количество точек растёт слишком быстро.

Full—Factorial



$N = 64$

Halton sequence

Определение

› Для заданной константы p определим ψ_p как

$$\psi_p(n) = \sum_{i=0}^{R(n)} \frac{a_i(n)}{p^{i+1}},$$

где a_i это числа из **записи** n **в системе счисления с основанием** p, а $R(n)$ обозначает максимальный индекс, для которого $a_i(n)$ не ноль.

› n-й элемент **последовательности Холтона** это

$$\mathbf{x}_n = (\psi_{p_1}(n), \ \ldots \ , \psi_{p_{d_{in}}}(n)),$$

где p_i это i-е простое число.

Halton sequence

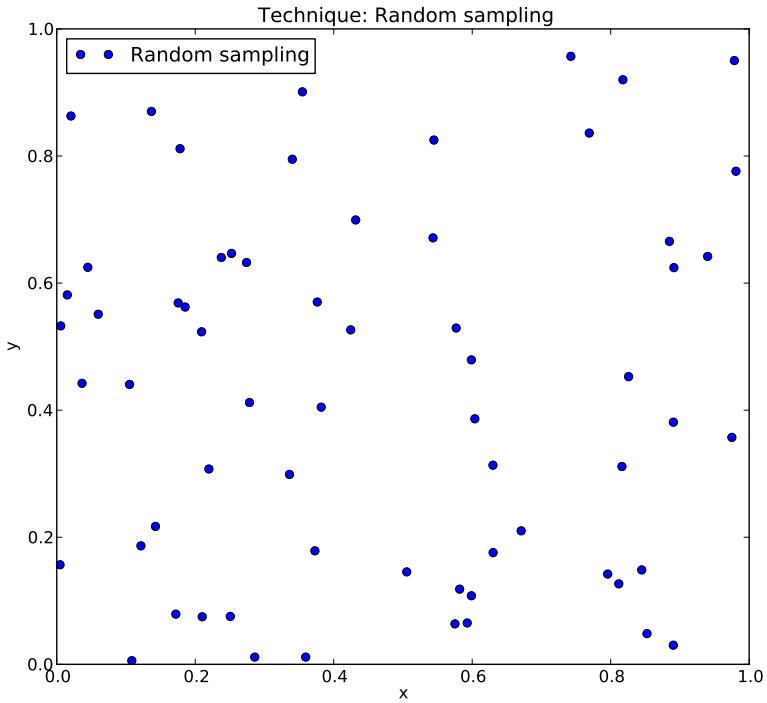
Преимущества

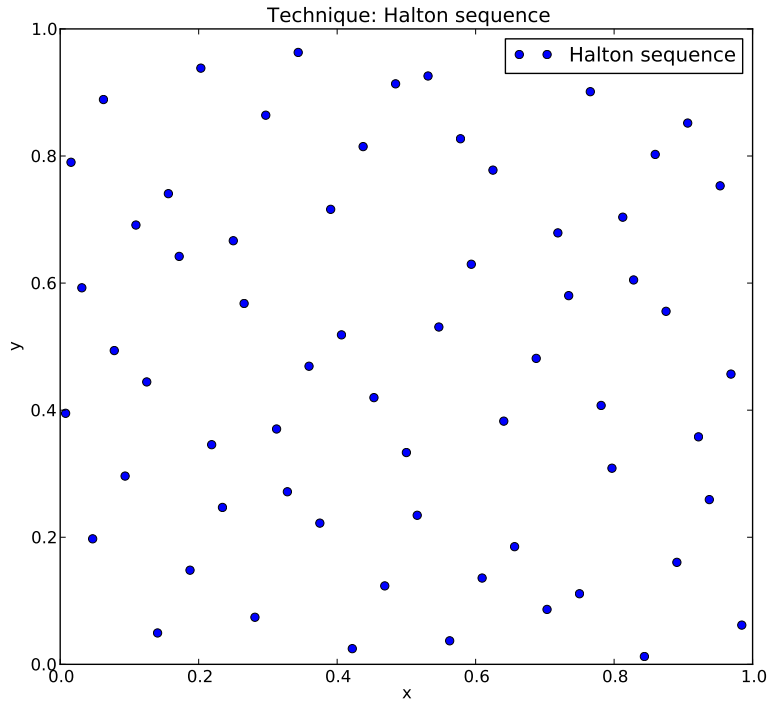
- Эффективность в малых размерностях.
- Всегда можно расширить добавлением новых точек.
- Быстрая генерация.

Недостатки

- Проекции с высокой размерностью ($d_{in} > 6$) сильно коррелированы.

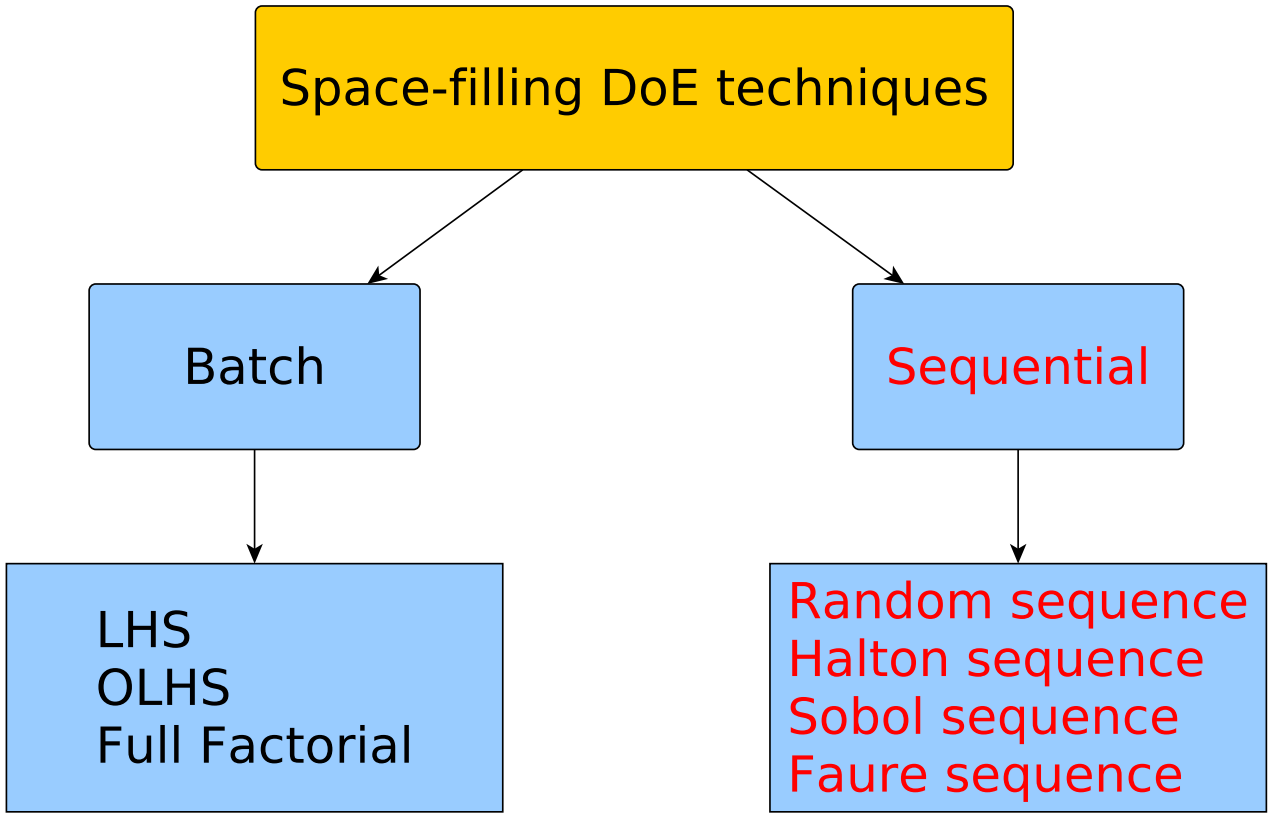
Halton sequence





$$d_{in} = 2\ ,\ N = 64$$

Техники Space-filling DoE



Свойства: Случайность или предопределённость

Некоторые техники DoE используют псевдо-случайные числа. Преимущество заключается в том, что результаты оказываются более разнообразными, но с другой стороны их труднее воспроизводить.

- **Включающие псевдо-случайность:** Random sampling, LHS, OLHS.
- **Полностью детерминированные:** Full Factorial, Halton sequence, Sobol sequence, Faure sequence.

Свойства: Добавление новых точек

Часть техник Space-filling поддерживает добавление новых точек к сгенерированному дизайну, а остальные не поддерживают из-за своих особенностей.

- **Поддерживающие добавление точек:** Random sampling, Halton sequence, Sobol sequence, Faure sequence. Эти техники могут быть использованы для построения дизайна от массива точек, так и в режиме последовательного добавления.
- **Сгенерированный дизайн не может быть расширен:** Full Factorial, LHS, OLHS. Эти техники позволяют только использование всех точек сразу.

Поверхность отклика

RSM

Определение

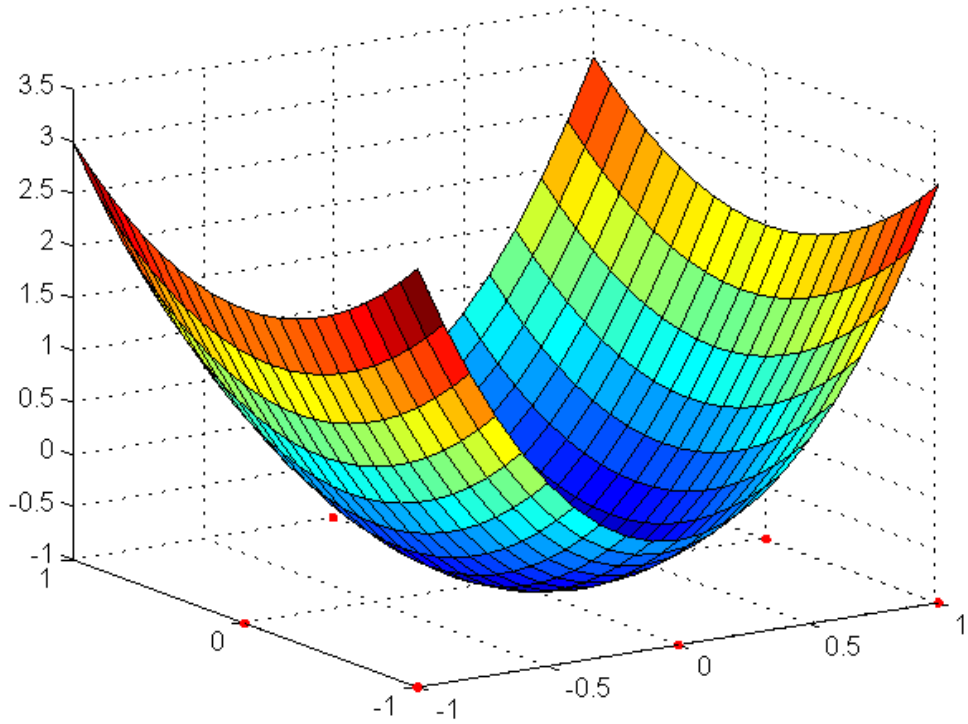
Response Surface Model (RSM):

$$\hat{f}(\mathbf{x}) = \alpha_0 + \sum_{i=1}^{d_{in}} \alpha_i x_i + \sum_{i,j=1,\, i \le j}^{d_{in}} \beta_{ij} x_i x_j \longleftrightarrow$$

$$\mathbf{x} = (x_1, \ldots, x_{d_{in}}) \in \mathbb{X}.$$

Параметры RSM настраиваются по обучающей выборке $D = (X, Y = f(X))$ с заранее заданным $X = (\mathbf{x}_i)_{i=1}^{N}$.

RSM



Модели RSM:

- линейная (все $\beta_{ij} = 0$)
- RSM с перекрёстными членами ($\beta_{ii} = 0$ для всех i)
- квадратичная

Оптимальные Дизайны RSM: введение

Response Surface Model (RSM):

$$\hat{f}(\mathbf{x}) = \alpha_0 + \sum_{i=1}^{d_{in}} \alpha_i x_i + \sum_{i,j=1,\, i\leq j}^{d_{in}} \beta_{ij} x_i x_j \longleftrightarrow$$

$$\mathbf{x} = (x_1, \ldots, x_{d_{in}}) \in \mathbb{X}.$$

Определение

Оптимальный Дизайн RSM это оптимальный $X = (\mathbf{x}_i)_{i=1}^N$, минимизирующий:

1. дисперсию оценки параметров RSM,
2. дисперсию прогноза модели,
3. другие оценки.

Оптимальные Дизайны для RSM

Оптимальные Дизайны минимизируют для RSM одну из величин:

— дисперсию оценки параметров RSM (**D-optimality**),

или

— дисперсию прогноза модели. (**IV-optimality**).

Преимущества

Оптимальные дизайны **уменьшают стоимость эксперимента** посредством того, что статистическая модель может быть оценена через меньшее число запусков.

Критерий оптимальности: Обозначения

- RSM может быть записан в виде

$$\hat{f}(\mathbf{x}) = \psi(\mathbf{x})\mathbf{c},$$

где $\mathbf{c} = (\mathbf{a}, \mathbf{b})$ и ψ — соответствующее отображение.

Например
$\psi(\mathbf{x}) = (1,\ x_1, \ldots, x_{d_{in}},\ x_1x_2,\ x_1x_3, \ldots, x_{d_{in}-1}x_{d_{in}})$ для RSM с `перекрёстными членами`.

- $\psi(X)$ будет обозначать $(\psi(\mathbf{x}_1), \ldots, \psi(\mathbf{x}_N))$.

Критерий оптимальности: Мотивация

- Предположим, что шум нормальный: $\mathbf{c} \sim \mathcal{N}(\hat{\mathbf{c}}, \mathrm{Cov}(\hat{\mathbf{c}}|X))$.
- Метод наименьших квадратов даёт оценку ковариации $\hat{\mathbf{c}}$.

$$\mathrm{Cov}(\hat{\mathbf{c}}|X) \sim \left(\psi(X)^T\psi(X)\right)^{-1}.$$

- Аналогично для прогноза ответа $f(\mathbf{x}_0) \sim \mathcal{N}(\hat{f}(\mathbf{x}_0), \mathrm{Var}(\hat{f}|X))$.
- Дисперсия прогноза $\hat{f}(\mathbf{x}_0) = \psi(\mathbf{x}_0)\hat{\mathbf{c}}$ в точке $\mathbf{x}_0 \in \mathbb{X}$.

$$\mathrm{Var}(\hat{f}|X) \sim \psi(\mathbf{x}_0)\left(\psi(X)^T\psi(X)\right)^{-1}\psi(\mathbf{x}_0)^T.$$

Критерий оптимальности: D-оптимальность

- Критерий D-оптимальности даёт такой дизайн, что детерминант должен быть минимален:

$$\det\left[\left(\psi(X)^T\psi(X)\right)^{-1}\right] \to \min_X .$$

- D-оптимальный дизайн минимизирует дисперсию оценки параметров.

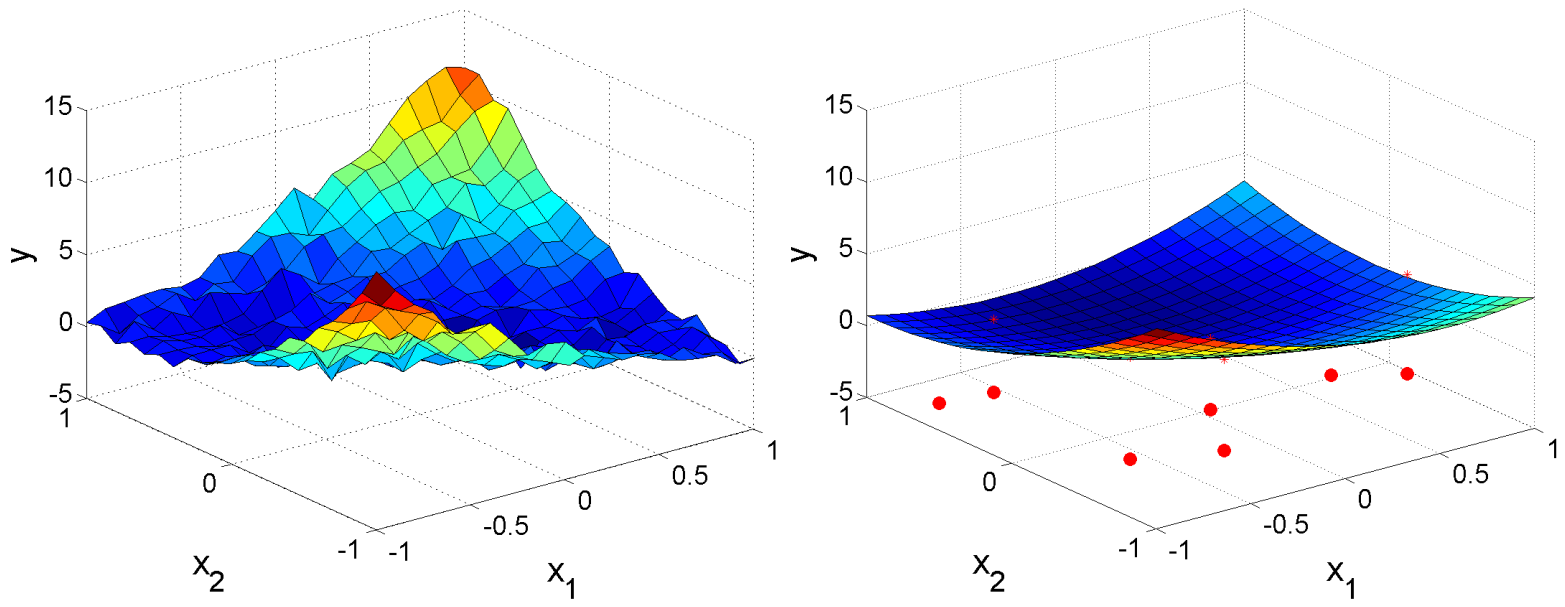
Критерий оптимальности: IV-оптимальность

› IV-оптимальный дизайн минимизирует дисперсия итогового прогноза модели:

$$\int_{\mathbb{X}} \psi(\mathbf{x}) \left(\psi(X)^T \psi(X)\right)^{-1} \psi(\mathbf{x})^T d\mathbf{x} \to \min_X .$$

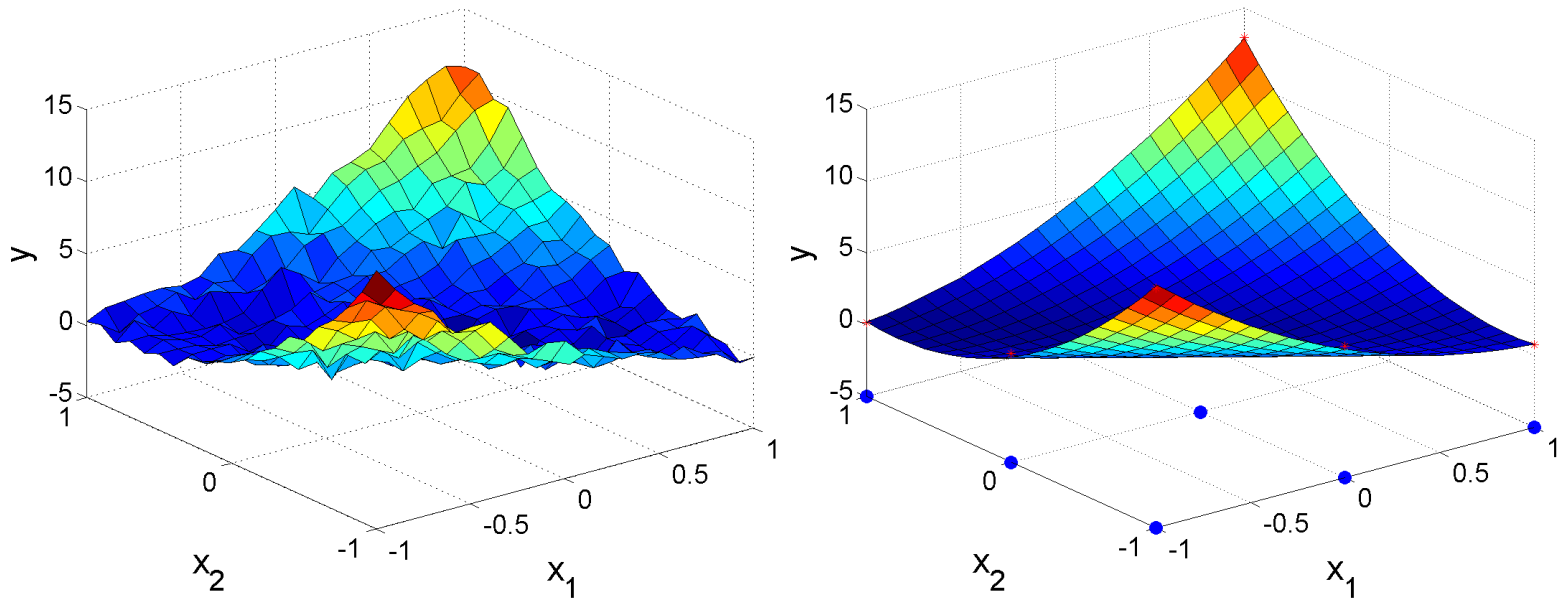
RSM: Пример 1

Настоящая функция ответа с шумом(слева) и RSM обученный со случайным DoE:



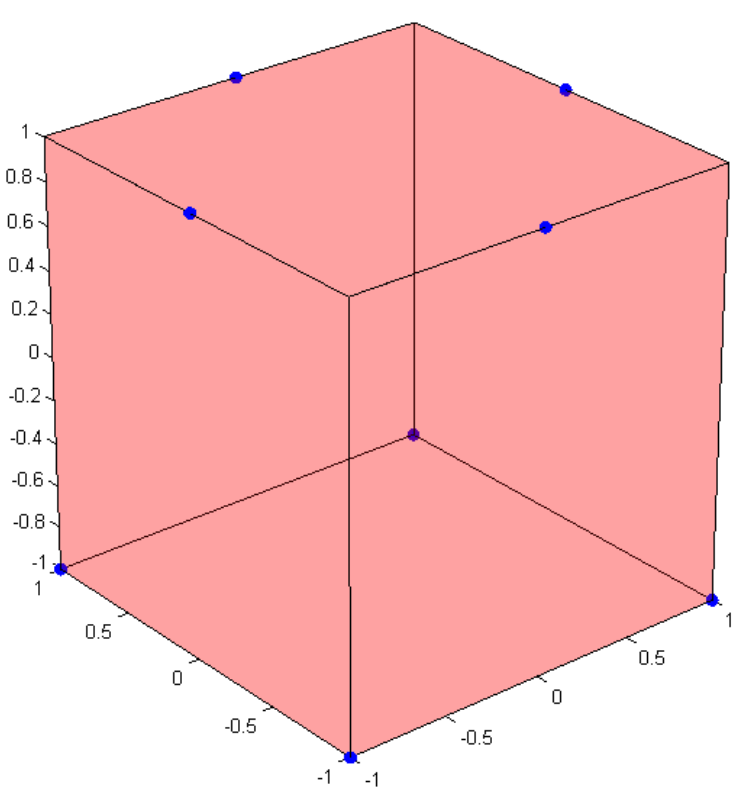
RSM: Пример 1

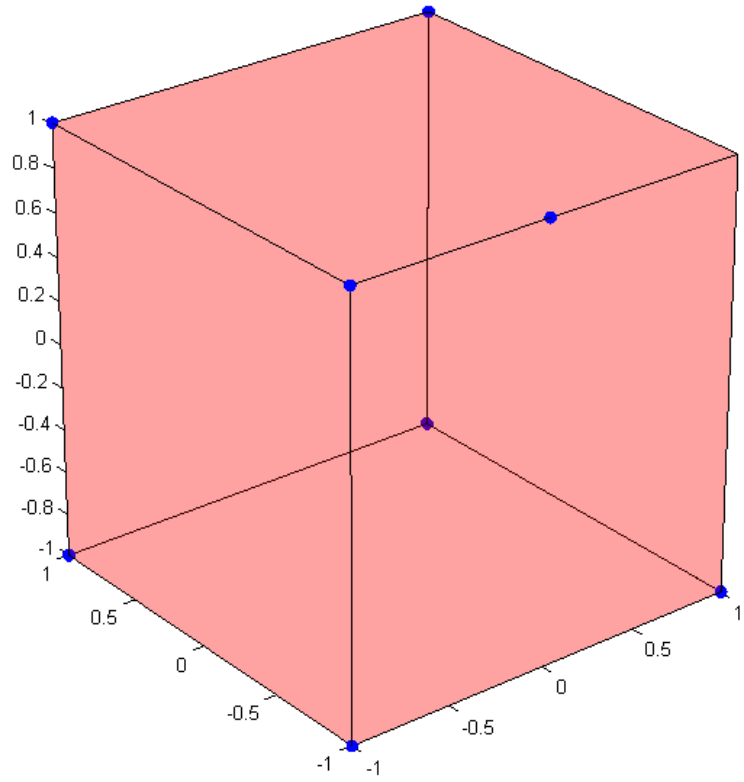
Настоящая зависимость(слева) и RSM обученный на D-оптимальном дизайне:



RSM: Пример 2

Оптимальный Дизайн для $\mathbb{X} \subset \mathbb{R}^3$:





RSM и категориальные переменные

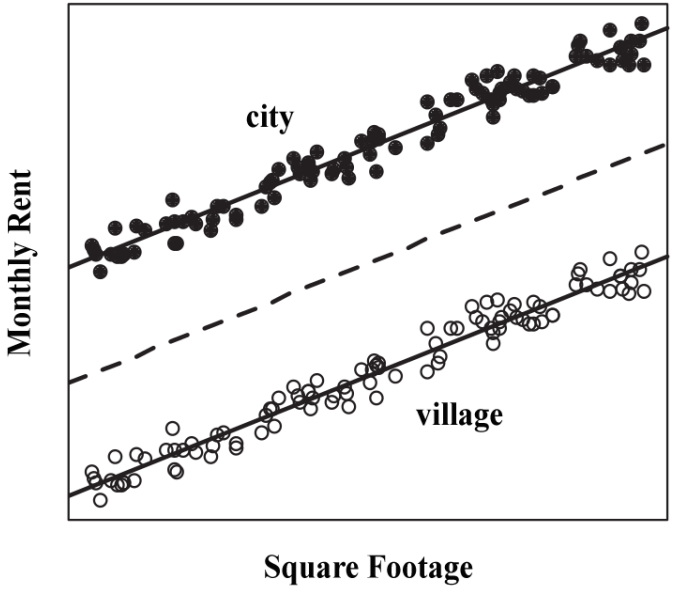
Определение

Категориальные переменные принимают одно из конечного числа значений. Они не имеют численного смысла или порядка.

Пример

Например, “цвет” (черный, зелёный, красный), “пол” (мужской, женский) и т.д.

Рисунок: линейная регрессия между месячной платой, типом жилья и квадратной площадью.



Категориальные переменные в RSM

Пусть $x \in \{v_1, \ldots, v_m\}$

1. Для каждого значения v_i строить свою модель RSM.
2. Dummy-variables. Заменим x на $y = (y_1, \ldots, y_m)$, где

$$y_i = \begin{cases} 1, \text{ если } x = v_i \\ 0 \text{ в противном случае.} \end{cases}$$

Критерий оптимальности: Алгоритм оптимизации

Оптимизационная процедура для поиска D- и IV- оптимальных дизайнов опирается на алгоритм Федорова:

Алгоритм

1. Установить число уровней для каждой переменной.
2. Сгенерировать полный факторный дизайн состоящий из всех комбинаций численных значений уровней. Это множество называется **множеством кандидатов** для Оптимального Дизайна.

Критерий оптимальности: Алгоритм оптимизации

Продолжение алгоритма

3. Взять необходимое число точек из **множества кандидатов** случайным образом и посчитать для них оптимальный критерий.

4. Продолжить оптимизацию добавлением и исключением точек из дизайна с целью минимизировать значение функционала. Новые точки берутся из **множества кандидатов**.

DoE для RSM

Преимущества

- Даёт наилучшие возможные оценки для Response Surface Models.
- Требует меньшее число запусков эксперимента, чем классические дизайны с той же точностью.
- Позволяет использовать категориальные переменные.

Недостатки

- Эффективность доказана только для Response Surface Model с априори заданной структурой.

Адаптивный Дизайн

Адаптивный Дизайн

Определение

Адаптивный DoE — это метод, который итеративно добавляет точки к обучающей выборке, минимизируя ошибку модели.

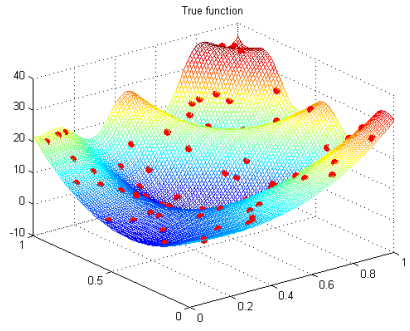


Рис.: Неизвестная функция

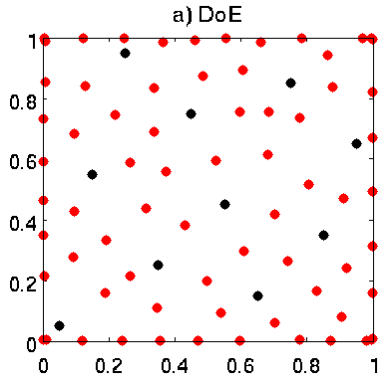


Рис.: Изначальные и добавленные точки

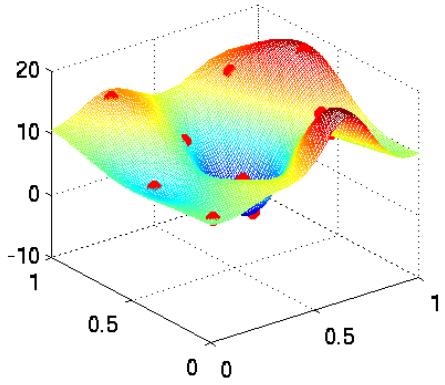


Рис.: Изначальная модель

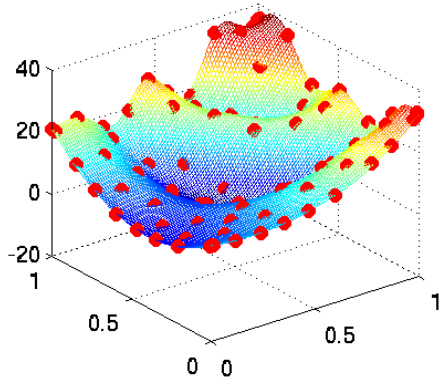
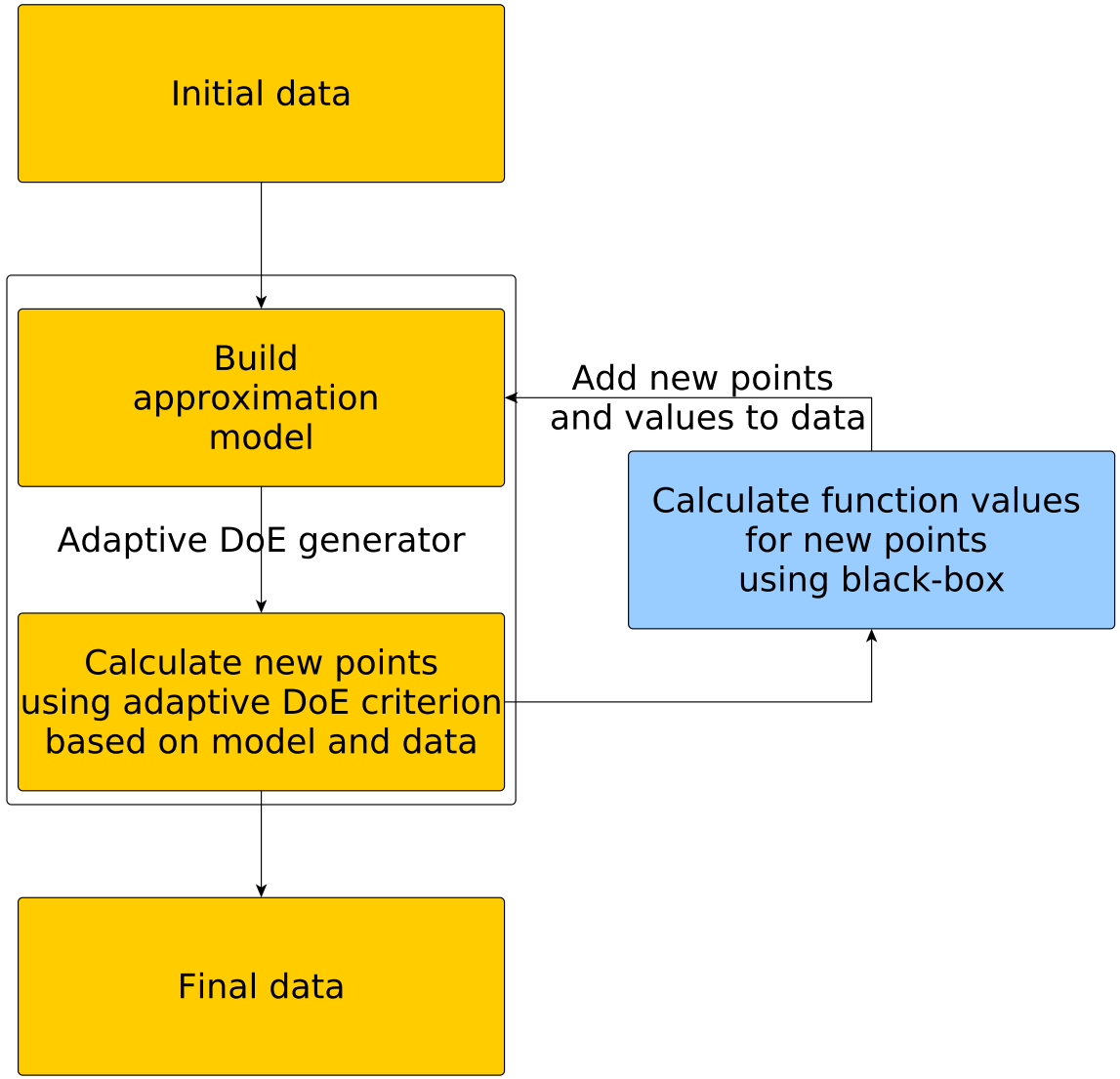


Рис.: Конечная модель

Адаптивный Дизайн

Адаптивный Дизайн эксперимента позволяет контролировать процесс суррогатного моделирования при помощи выборочного семплирования, улучшая тем самым качество аппроксимации.
1

Процесс Адаптивного DoE



Адаптивный DoE
Причины такого подхода:

1. Аппроксимация содержит в себе информацию о поведении функции.

   ⇓

   Аппроксимация может подсказать, как выбрать новый элемент, чтобы увеличить качество больше всего.

2. Адаптивное обновление обучающей выборки даёт гибкие возможности для изменения числа точек.

   ⇓

   Если желаемое качество аппроксимации достигнуто, процесс DoE может быть закончен.

Регрессия гауссовского процесса

Регрессия

Стандартная линейная регрессия

Множественная регрессия

Прогнозирование

Выбор модели

Дизайн эксперимента

Заполнение пространства

Random sampling

Семплирование латинскими гиперкубами

Полный факторный дизайн

Halton sequence

Свойства

Поверхность отклика

Оптимальные Дизайны RSM

Примеры

RSM и категориальные переменные

DoE для RSM

Адаптивный Дизайн

Техника Адаптивного DoE

Регрессия гауссовского процесса

Предположения

Выход суррогатной модели

Критерий сэмплирования

Integrated MSE Gain-Maximum Variance

Поверхность критерия максимальной дисперсии

MaxMin

Пример

Предположения

Пусть задано случайное поле $f(\mathbf{x}, \omega)$, $\omega \in \Omega$ — случайное событие в Ω. Предполагается, что для произвольных точек из области $\mathbb{X}$ существуют первый и второй моменты:

$$M(\mathbf{x}) = \mathbb{E}f(\mathbf{x}),$$

$$K(\mathbf{x}_1, \mathbf{x}_2) = \mathbb{E}(f(\mathbf{x}_1) - \mathbb{E}f(\mathbf{x}_1))(f(\mathbf{x}_2) - \mathbb{E}f(\mathbf{x}_2)),$$

а также условное математическое ожидание

$$\mathbb{E}(f(\mathbf{x})|f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_l)).$$

Предположим также, что случайное поле — гауссовское. Для такого поля совместное распределение $f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_l)$ — нормальное и, следовательно, определяется математическим ожиданием и ковариационной функцией.

Вид ковариационной функции

Пусть зависимость $y(\mathbf{x})$ порождена моделью:

$$y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon(\mathbf{x}),$$

где $f(\mathbf{x})$ — некоторая реализация случайного гауссовского поля, а $\varepsilon(\mathbf{x}) \sim \mathcal{N}(0, \tilde{\sigma}^2)$ — гауссовский белый шум.
Предположим, что ковариационная функция $k(\mathbf{x}, \mathbf{x}')$ гауссовского поля $f(\mathbf{x})$ принадлежит некоторому параметрическому семейству

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 k(\mathbf{x}, \mathbf{x}'|\Theta),$$

где Θ — некоторый набор параметров, σ^2 - параметр масштаба ковариационной функции.
Тогда ковариационная функция процесса $y(\mathbf{x})$ имеет вид:

$$K(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') + \tilde{\sigma}^2 \delta(\mathbf{x}, \mathbf{x}'),$$

где $\delta(\mathbf{x}, \mathbf{x}')$ — символ Кронекера.

Регрессия гауссовского процесса

- Предположим, что $f(\mathbf{x})$ является реализацией гауссовского процесса (GP) со средним $\mu(\mathbf{x})$ и ковариационной функцией $k(\mathbf{x}, \mathbf{x}')$.
- GP это стохастический процесс, все конечные сечения которого $f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_l)$— гауссовские.
- Гауссовский процесс полностью определяется своим средним и ковариационной функцией.
- Предположим, что $\mu(\mathbf{x}) \equiv 0$ и ковариационная функция $k(\mathbf{x}, \mathbf{x}')$ известна.

Регрессия гауссовского процесса

Предположим, данные зашумлены: $y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon(\mathbf{x})$, где $\varepsilon(\mathbf{x})$ — гауссовский белый шум,
то есть $\forall \mathbf{x}_1, \ldots, \mathbf{x}_l \in \mathbb{X} \subseteq \mathbb{R}^{d_{in}} : \varepsilon(\mathbf{x}_1), \ldots, \varepsilon(\mathbf{x}_l)$ удовлетворяют свойствам:

- Независимо и одинаково распределены.
- Распределение нормально.
- С нулевым средним и дисперсией $\tilde{\sigma}^2$.

Регрессия гауссовского процесса

$D = (X, \mathbf{y}) = \{(\mathbf{x}_i, y_i = f(\mathbf{x}_i) + \varepsilon(\mathbf{x}_i))\}_{i=1}^N$ — обучающее множество.

Здесь и далее для простоты возьмём $d_{out} = 1$, то есть $Y = \mathbf{y} \in \mathbb{R}^1$.

Апостериорное распределение $f(\mathbf{x})$:

$$\mathrm{Law}(f(\mathbf{x})|D) = \mathcal{N}(\hat{f}(\mathbf{x}), \hat{\sigma}^2(\mathbf{x})).$$

Выход суррогатной модели

- Апостериорное среднее, используемое как выход суррогатной модели $\hat{f}(\mathbf{x})$, имеет вид:

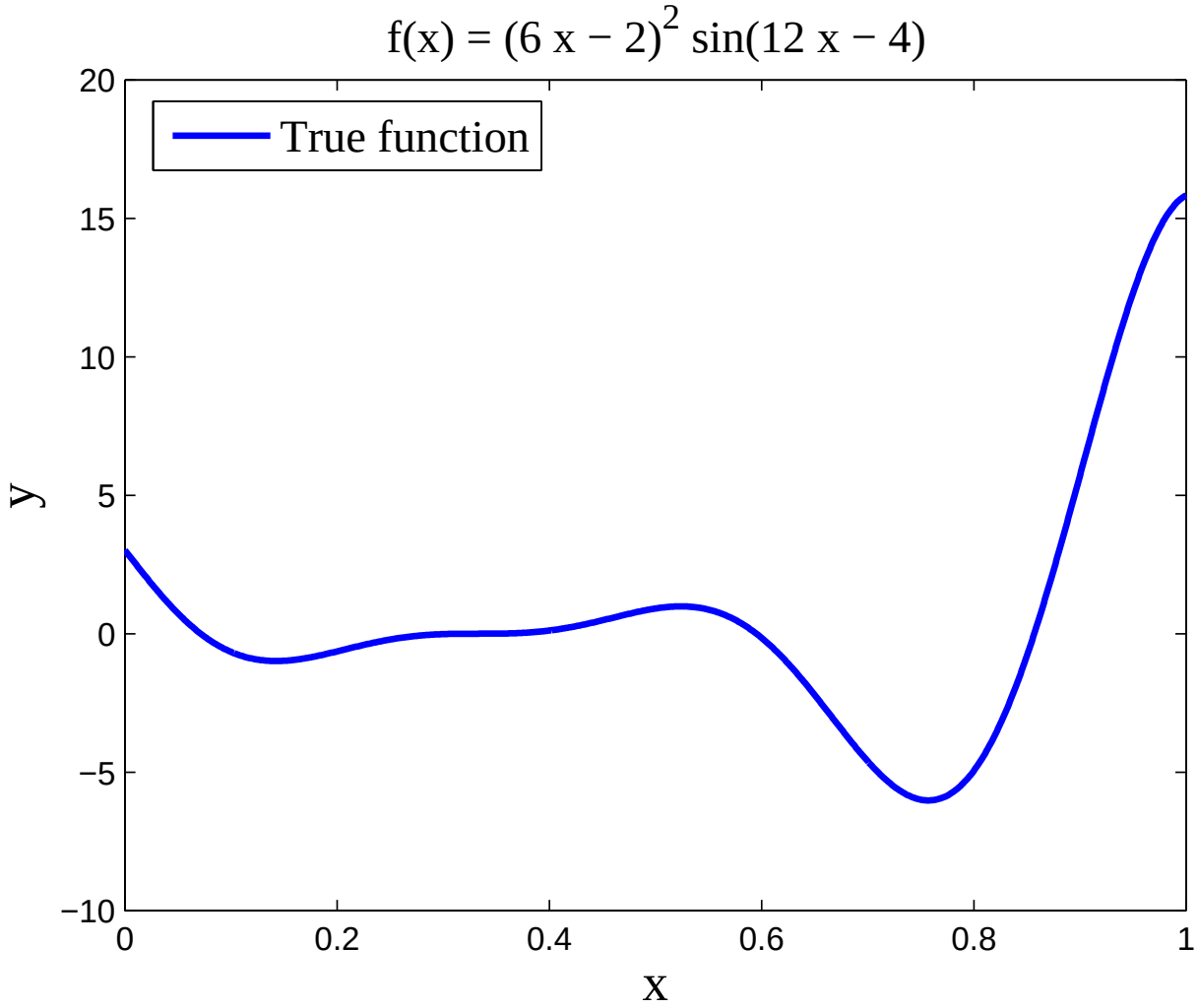
$$\hat{f}(\mathbf{x}) = \mathbf{k}(\mathbf{x})^T(\mathrm{K} + \tilde{\sigma}^2\mathrm{I})^{-1}\mathbf{y},$$

  где $\mathrm{K} = \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^N$, $\mathbf{k}(\mathbf{x}) = \{k(\mathbf{x}, \mathbf{x}_i)\}_{i=1}^N$.

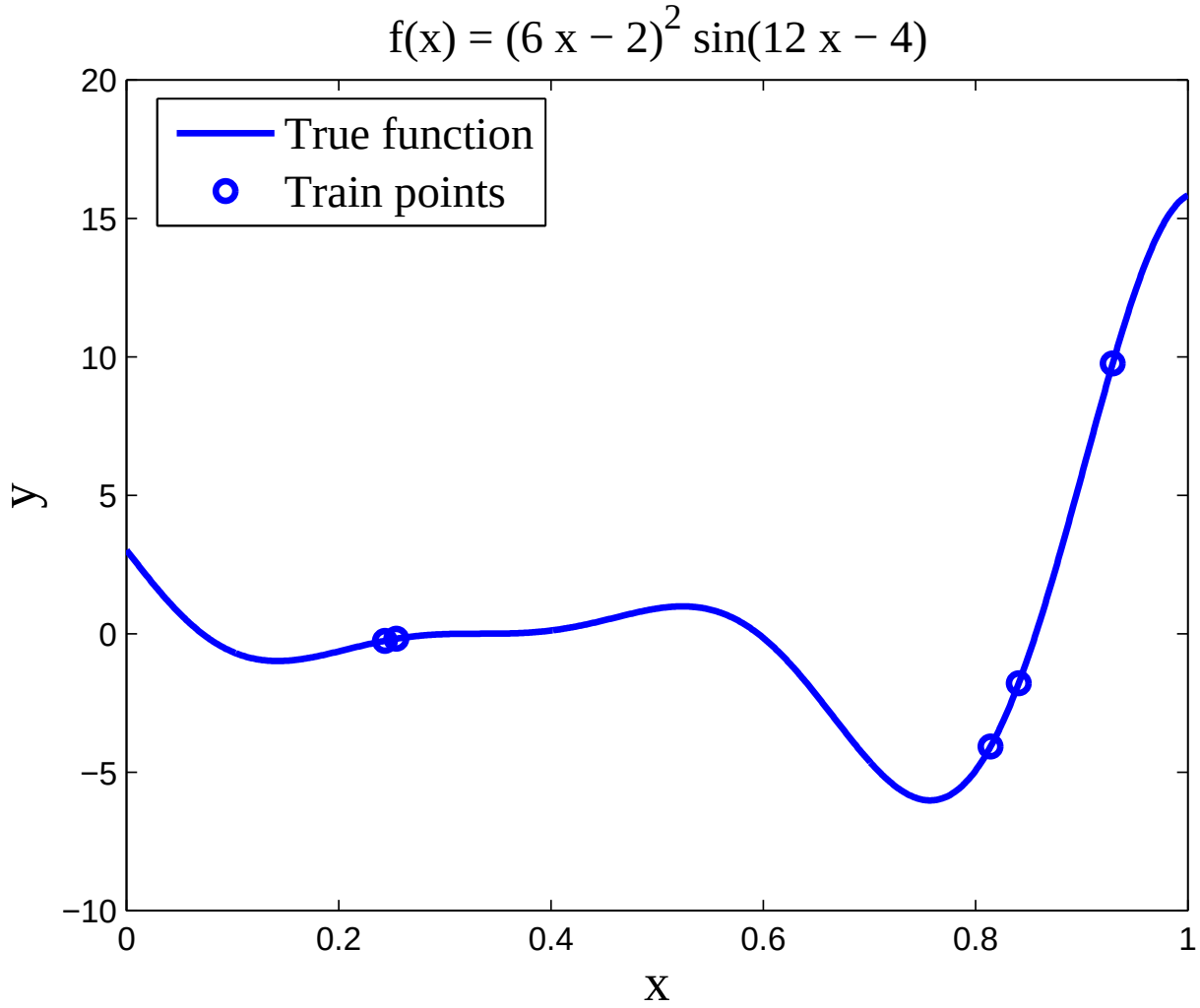
- Апостериорная ковариация, используемая как оценка точности, имеет вид:

$$\hat{\sigma}^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^T(\mathrm{K} + \tilde{\sigma}^2\mathrm{I})^{-1}\mathbf{k}(\mathbf{x}).$$

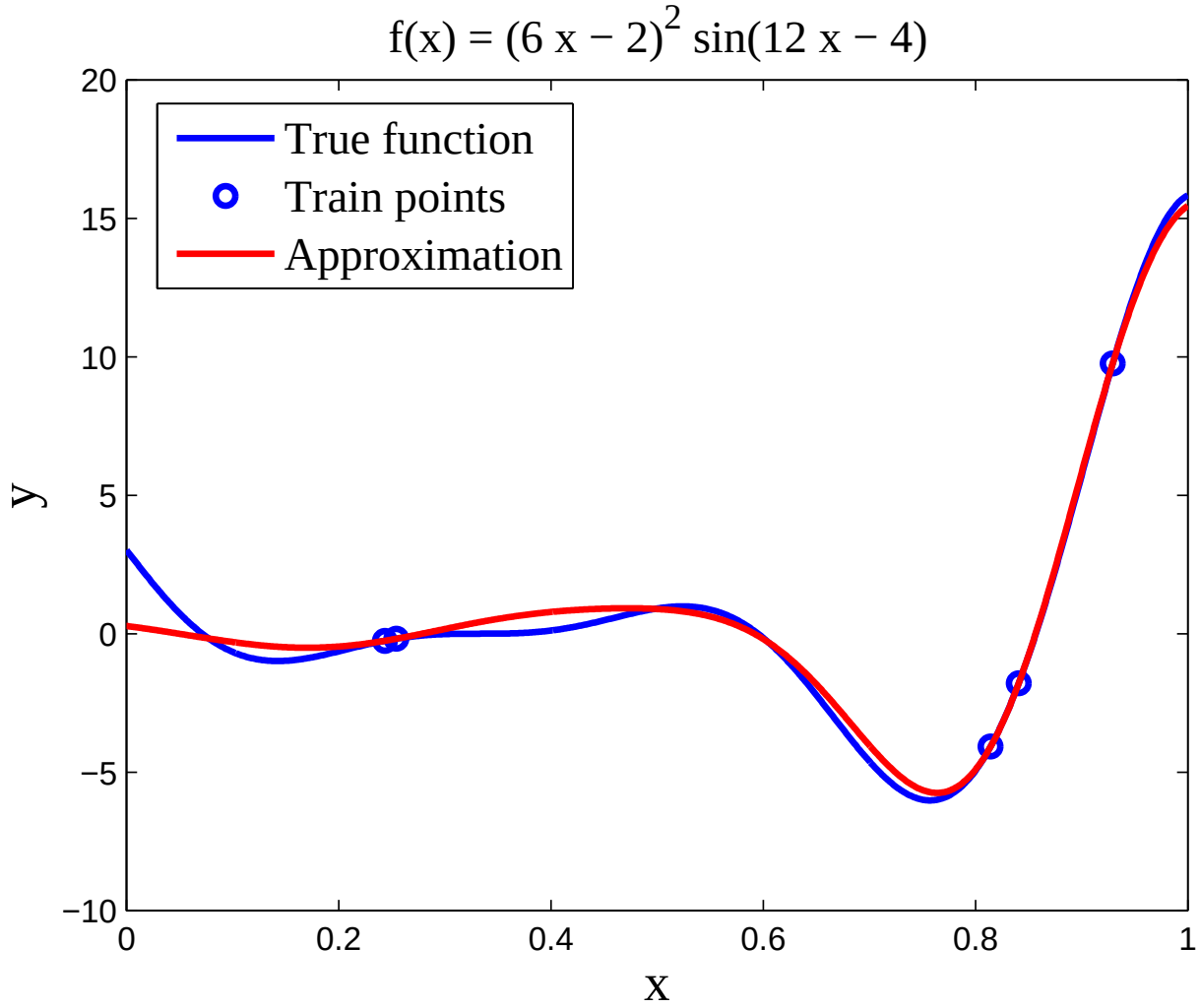
Выход суррогатной модели



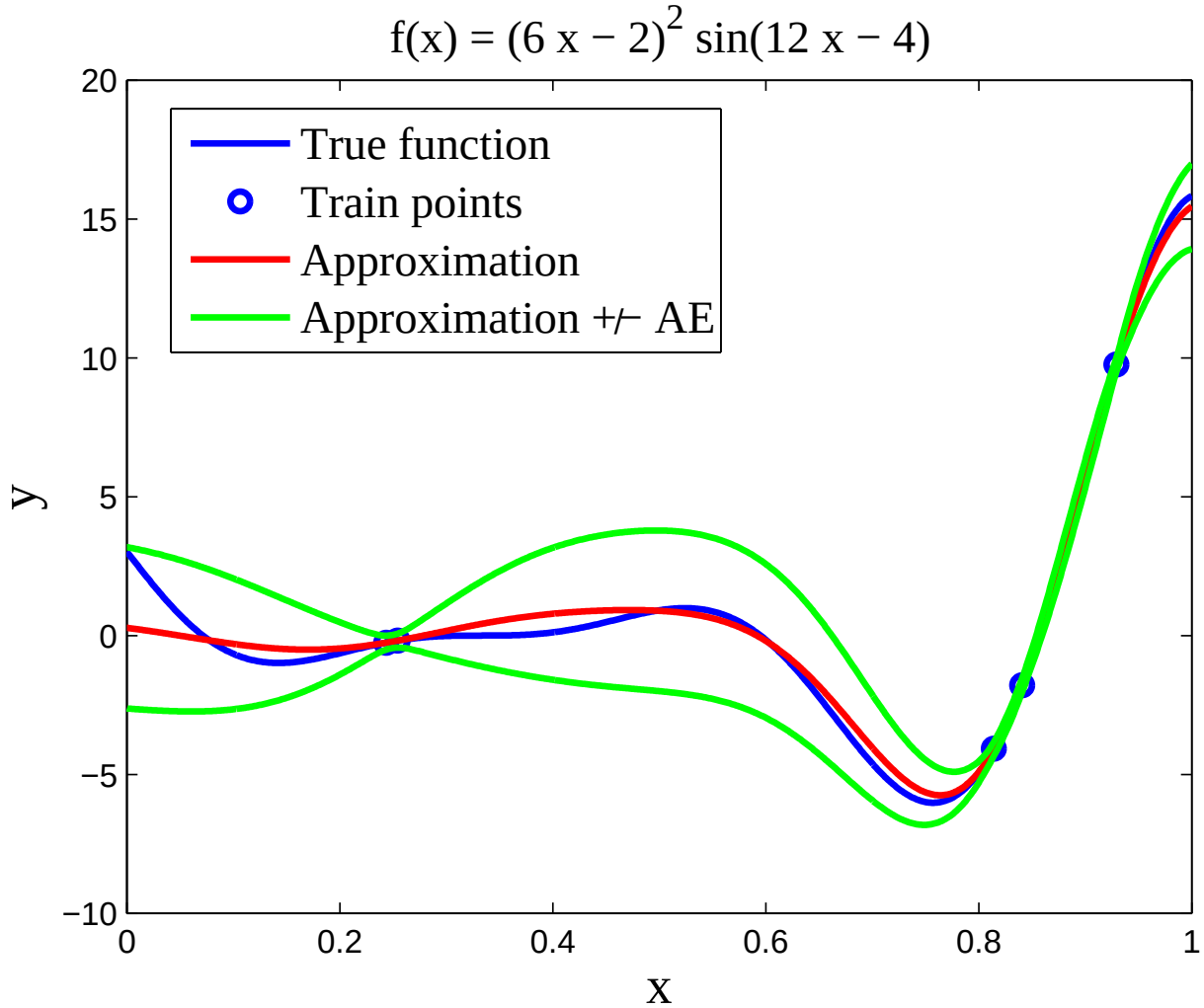
Выход суррогатной модели



Выход суррогатной модели



Выход суррогатной модели



Ковариационная функция

В реальных задачах ковариационная функция обычно неизвестна, и нужно её оценить по данным.

Для оценки ковариации используются два семейства функций:

Определение

- Взвешенное Евклидово расстояние:

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\sum_{i=1}^{d_{in}} \theta_i^2 (x_i - x_i')^s\right), s \in [1, 2],$$

- Расстояние Махаланобиса:

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\left(\mathbf{x} - \mathbf{x}'\right)^T \Theta \left(\mathbf{x} - \mathbf{x}'\right)\right),$$

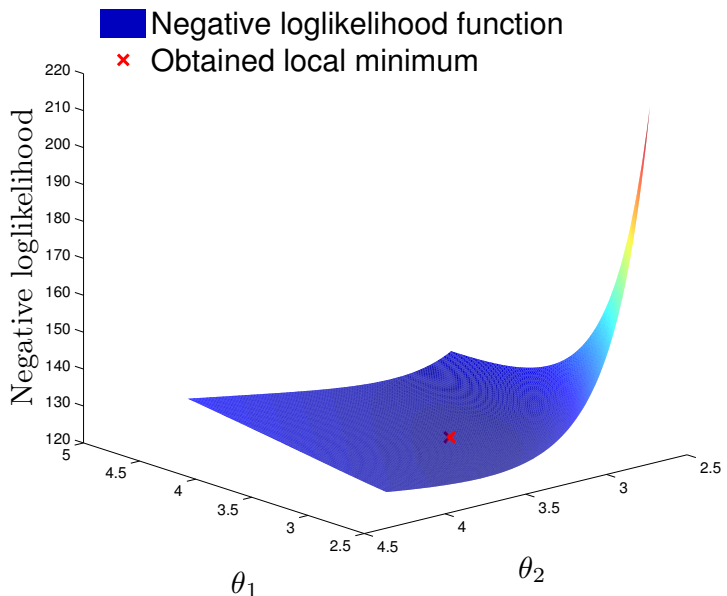
где Θ это $d_{in} \times d_{in}$ матрица.

Оценивание параметров

Параметры ковариационной функции $k(\mathbf{x}, \mathbf{x}')$ оцениваются при помощи максимизации правдоподобия:

$$-\frac{1}{2}\ln|\mathrm{K}| - \frac{\mathbf{y}^T\mathrm{K}^{-1}\mathbf{y}}{2} \to \max_{\mathbf{a}},$$

где $\mathbf{a} = (\theta_1, \ldots, \theta_{d_{in}}; \sigma)$ или $\mathbf{a} = (\Theta; \sigma)$ в зависимости от типа используемой ковариационной функции.



Критерий сэмплирования

Регрессия

Стандартная линейная регрессия

Множественная регрессия

Прогнозирование

Выбор модели

Дизайн эксперимента

Заполнение пространства

Random sampling

Семплирование латинскими гиперкубами

Полный факторный дизайн

Halton sequence

Свойства

Поверхность отклика

Оптимальные Дизайны RSM

Примеры

RSM и категориальные переменные

DoE для RSM

Адаптивный Дизайн

Техника Адаптивного DoE

Регрессия гауссовского процесса

Предположения

Выход суррогатной модели

Критерий сэмплирования

Integrated MSE Gain-Maximum Variance

Поверхность критерия максимальной дисперсии

MaxMin

Пример

Критерий сэмплирования

Процедура выбора новой точки для дизайна может быть сформулирована как задача максимизации некоторого критерия:

$$\mathbf{x}_{new} = \arg\max_{\mathbf{x}\in\mathbb{X}} \mathcal{I}(\mathbf{x}|D, \hat{f}, \hat{\sigma}^2),$$

где $\mathcal{I}(\mathbf{x}|D, \hat{f}, \hat{\sigma}^2)$ - это критерий выбора точек, основанный на обучающем множестве D, текущей аппроксимации $\hat{f}$ и текущей ошибке $\hat{\sigma}^2$.

Критерий максимальной дисперсии

Определение

Критерий максимальной дисперсии это

$$\mathcal{I}_{MV}(\mathbf{x}) = \hat{\sigma}^2(\mathbf{x}|X),$$

где $\hat{\sigma}^2(\mathbf{x}|X)$ - ошибка обученной на $D = (X, \mathbf{y})$ аппроксимации в точке $\mathbf{x}$.

Преимущества

- Легко вычислим.
- Включает информацию о поведении функции.

Недостатки

- Учитывается только локальное поведение.
- Склонна давать точки ближе к границе $\mathbb{X}$.

Минимизация средней ошибки предсказания

Определение

Оптимальный критерий для минимизация ожидаемой среднего квадрата ошибки аппроксимации на следующей итерации:

$$\mathcal{I}_{\rho_2}(\mathbf{x}) = \frac{1}{|\mathbb{X}|}\int_{\mathbb{X}}\big(\hat{\sigma}^2(\mathbf{v}|X) - \hat{\sigma}^2(\mathbf{v}|X \cup \mathbf{x})\big)\mathrm{d}\mathbf{v},$$

где

- $\hat{\sigma}^2(\mathbf{v}|X)$ - ошибка предсказания аппроксимации построенной на множестве $D = (X, \mathbf{y})$.
- $\hat{\sigma}^2(\mathbf{v}|X \cup \mathbf{x})$ - ошибка предсказания аппроксимации построенной на множестве $D^{ext} = D \cup (\mathbf{x}, y(\mathbf{x}))$.

Integrated MSE Gain-Maximum Variance

Определение

Integrated MSE Gain-Maximum Variance критерий — это

$$\mathcal{I}_{IGMV}(\mathbf{x}) = \mathcal{I}_{\rho_2}(\mathbf{x}) * \mathcal{I}_{MV}(\mathbf{x}).$$

Преимущества

- Включает информацию о поведении функции на всём пространстве дизайна.

Недостатки

- Относительно сложная вычислимость.

Взгляд на критерий ImseGain-MaxVar изнутри

Некоторые матричные вычисления позволяют выразить этот критерий следующим образом:

$$\mathcal{I}_{\rho_2}(\mathbf{x}) = \frac{1}{|\mathbb{X}|}\int_{\mathbb{X}} \frac{\hat{K}^2(\mathbf{x},\mathbf{v})}{\hat{\sigma}^2(\mathbf{x}|X)}\mathrm{d}\mathbf{v},$$

где $\hat{K}(\mathbf{x},\mathbf{v}) = K(\mathbf{x},\mathbf{v}) - \mathbf{k}(\mathbf{x})\mathrm{K}^{-1}\mathbf{k}(\mathbf{v})^T$ — апостериорная ковариция между значениями гауссовского процесса в точках $\mathbf{x}$ и $\mathbf{v}$.

Критерий может быть вычислительно нестабильным, если $\hat{\sigma}^2(\mathbf{x}|X)$ мало.

Взгляд на критерий ImseGain-MaxVar изнутри

Можно решить эту проблему, совместив критерии $\mathcal{I}_{\rho_2}(\mathbf{x})$ и $\mathcal{I}_{MV}(\mathbf{x})$ в один с меньшим числом недостатков:

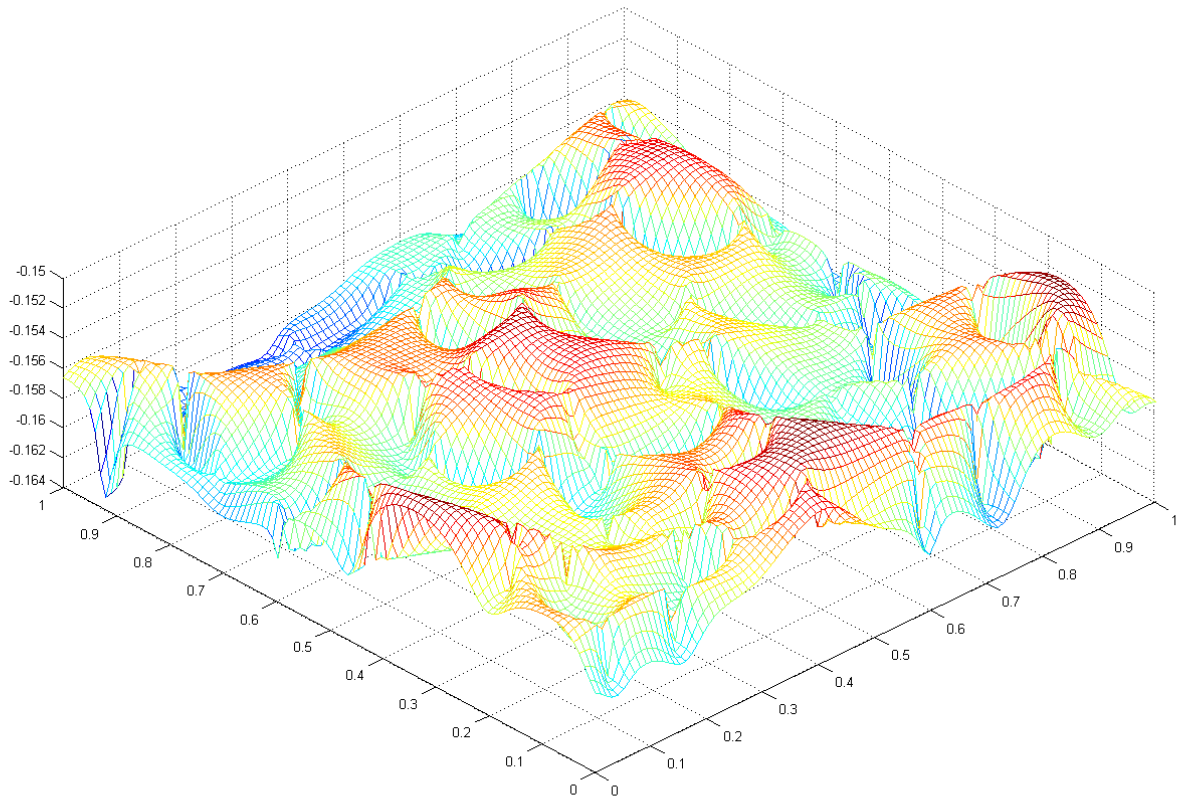
$$\mathcal{I}_{IGMV}(\mathbf{x}) = \mathcal{I}_{\rho_2}(\mathbf{x}) * \mathcal{I}_{MV}(\mathbf{x}) = \frac{1}{|\mathbb{X}|}\int_{\mathbb{X}} \hat{K}^2(\mathbf{x}, \mathbf{v})\mathrm{d}\mathbf{v}.$$

Преимущества

- Нет проблемных членов в знаменателе.
- По-прежнему ведёт к аппроксимации по всему пространству дизайна.

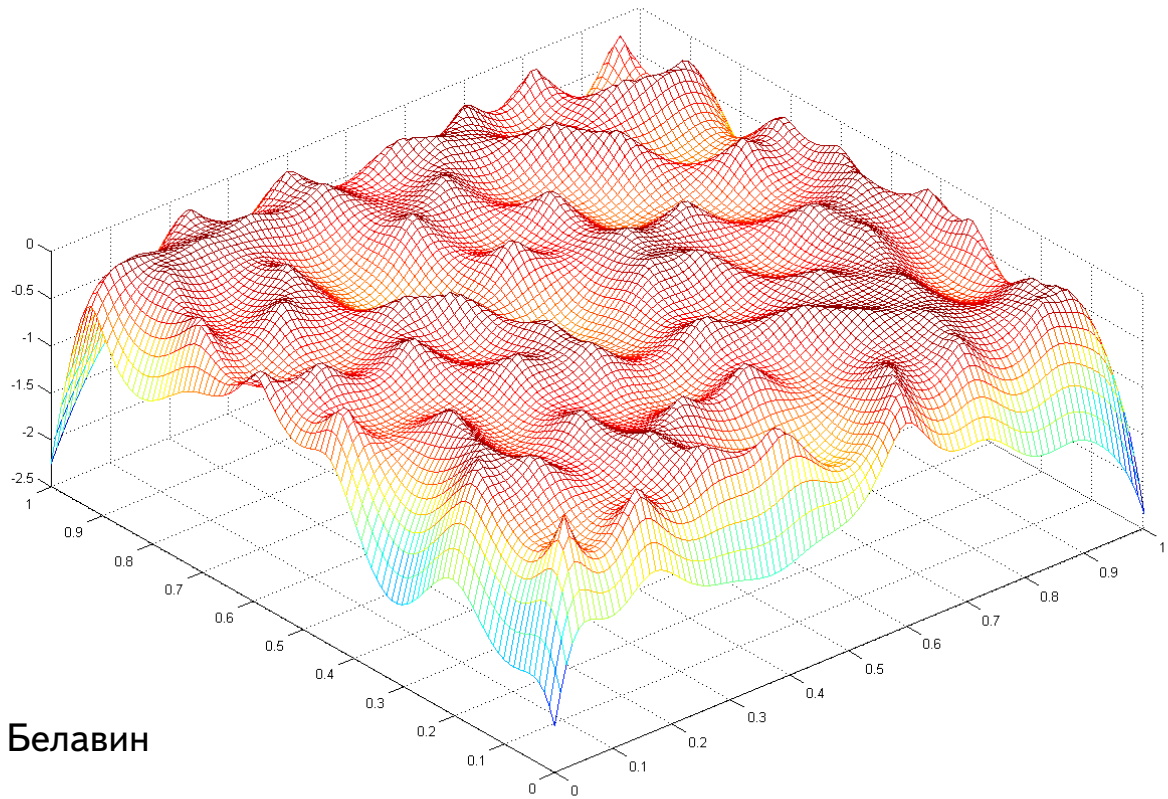
$\mathcal{I}_{\rho_2}$ поверхность

- Высокая мультимодальность.
- Узкие и не робастные локальные минимумы.



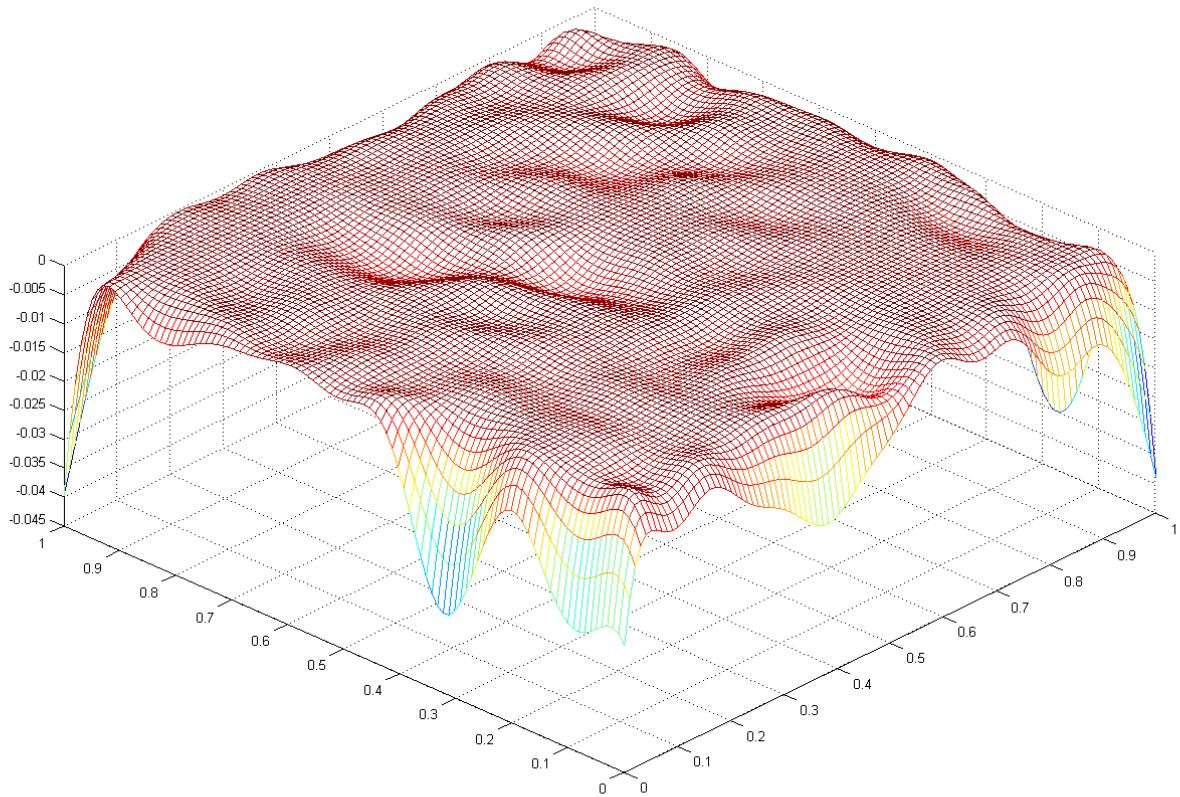
Поверхность критерия максимальной дисперсии

- Высокая мультимодальность.
- Поведение более регулярное.



Поверхность ImseGain-MaxVar

- Меньше локальных минимумов.
- Регулярность поведения.



MaxMin

Определение

MaxMin критерий это:

$$\mathcal{I}_{MM}(\mathbf{x}) = \min_{\mathbf{v} \in X} d^2(\mathbf{v}, \mathbf{x}),$$

где $d(\mathbf{v}, \mathbf{x})$ — это Евклидово расстояние между точками $\mathbf{v}$ и $\mathbf{x}$: ($d^2(\mathbf{v}, \mathbf{x}) = \sum_{i=1}^{d_{in}} (v_i - x_i)^2$).

Преимущества

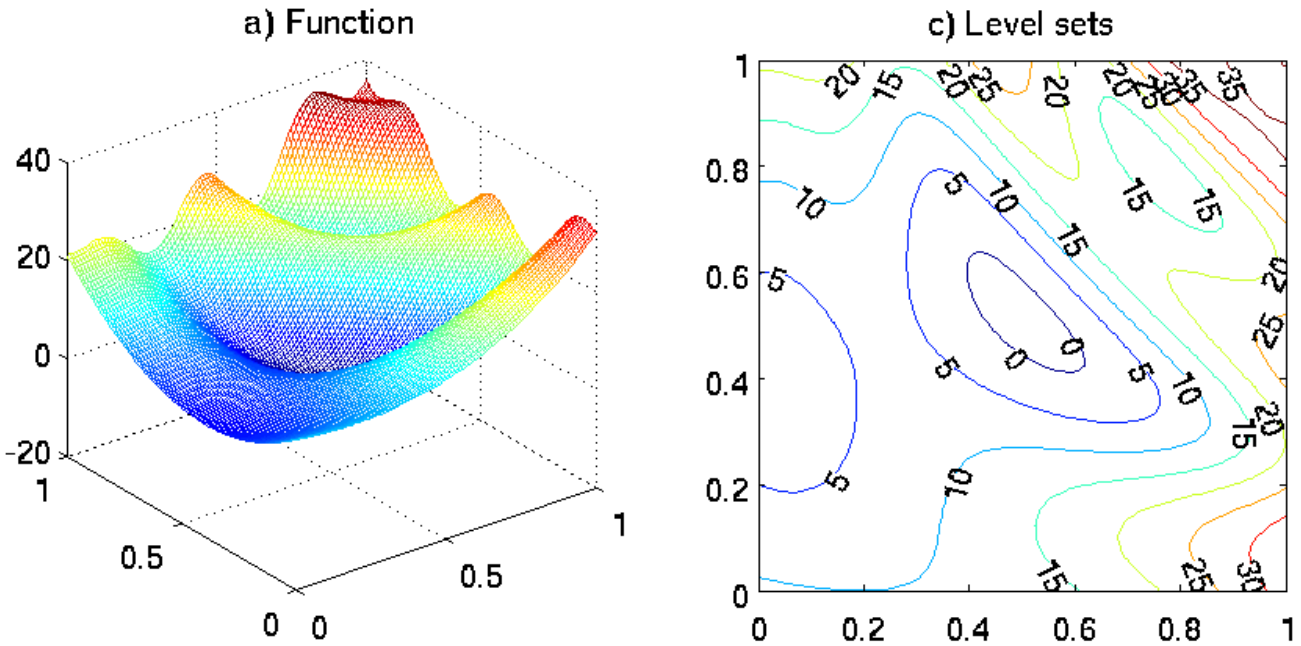
- Очень быстро вычисляется.
- Не опирается на тип аппроксимируемой модели.
- Берёт точки равномерно по пространству дизайна.

Недостатки

Пример: Неизвестная функция

Рассмотрим задачу аппроксимации следующей функции:

$$y = 2 + 0.25(x_2 - 5x_1^2)^2 + (1 - 5x_1)^2 + 2(2 - 5x_2)^2 + 7sin(2.5x_1)$$



Пример: начальная аппроксимация

Изначальный экспериментальный дизайн содержит 10 точек.
Строим аппроксимацию регрессией гауссовского процесса.

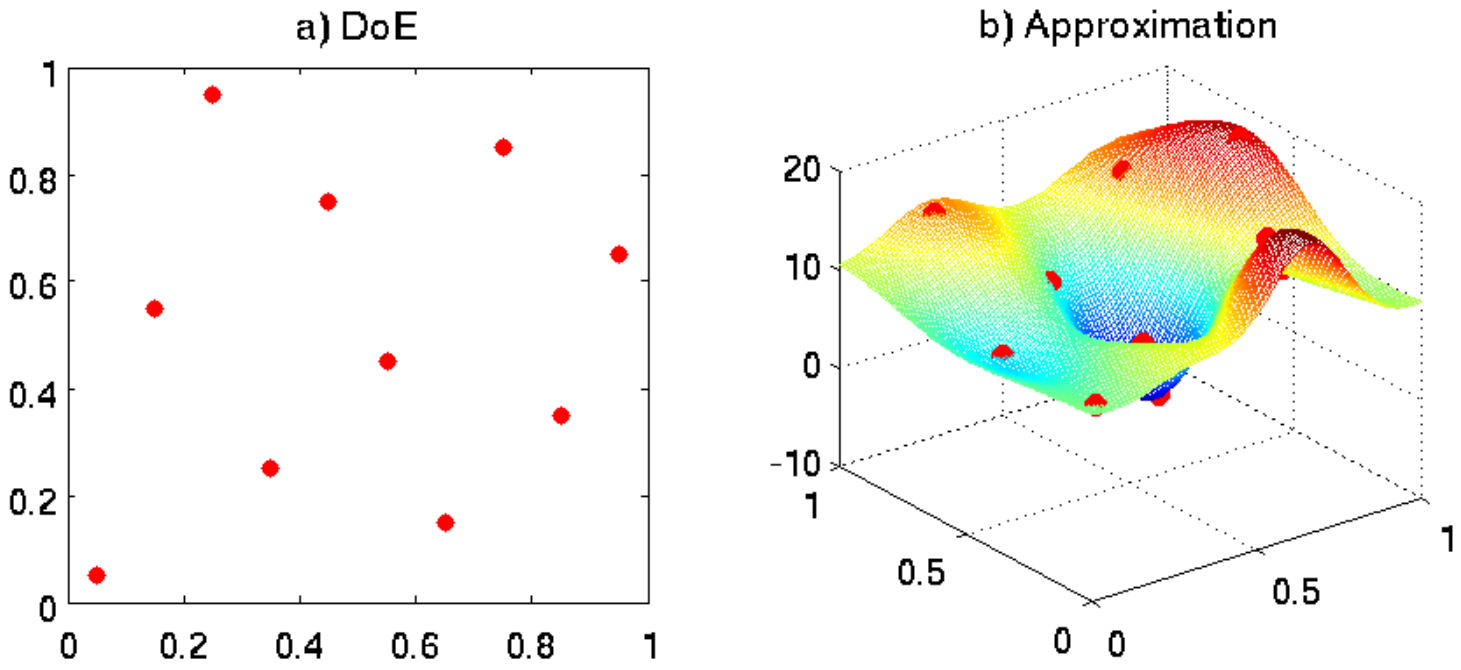
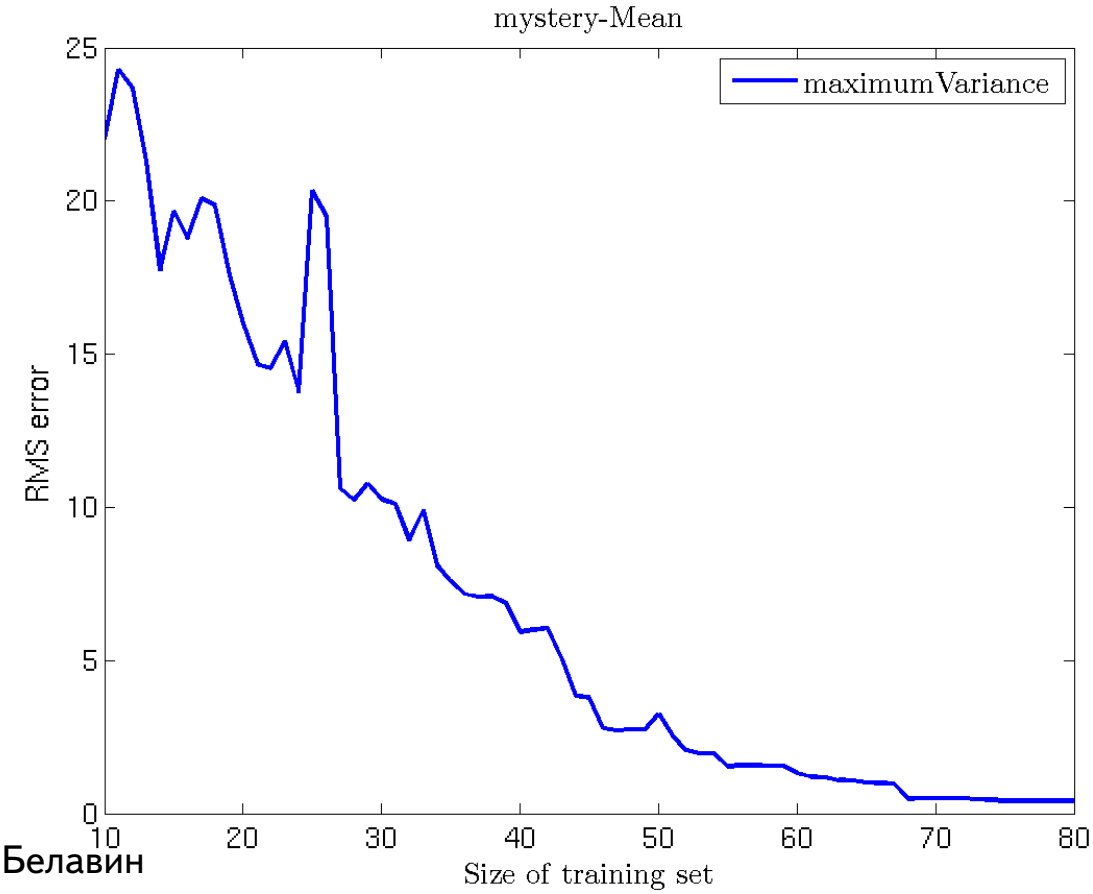


Рис.: Начальная обучающая выборка и аппроксимация

Пример: адаптивный процесс DoE

Добавлено 70 точек при помощи критерия Максимальной Дисперсии.

Изменение ошибки показано на рисунке ниже.



Пример: финальная аппроксимация

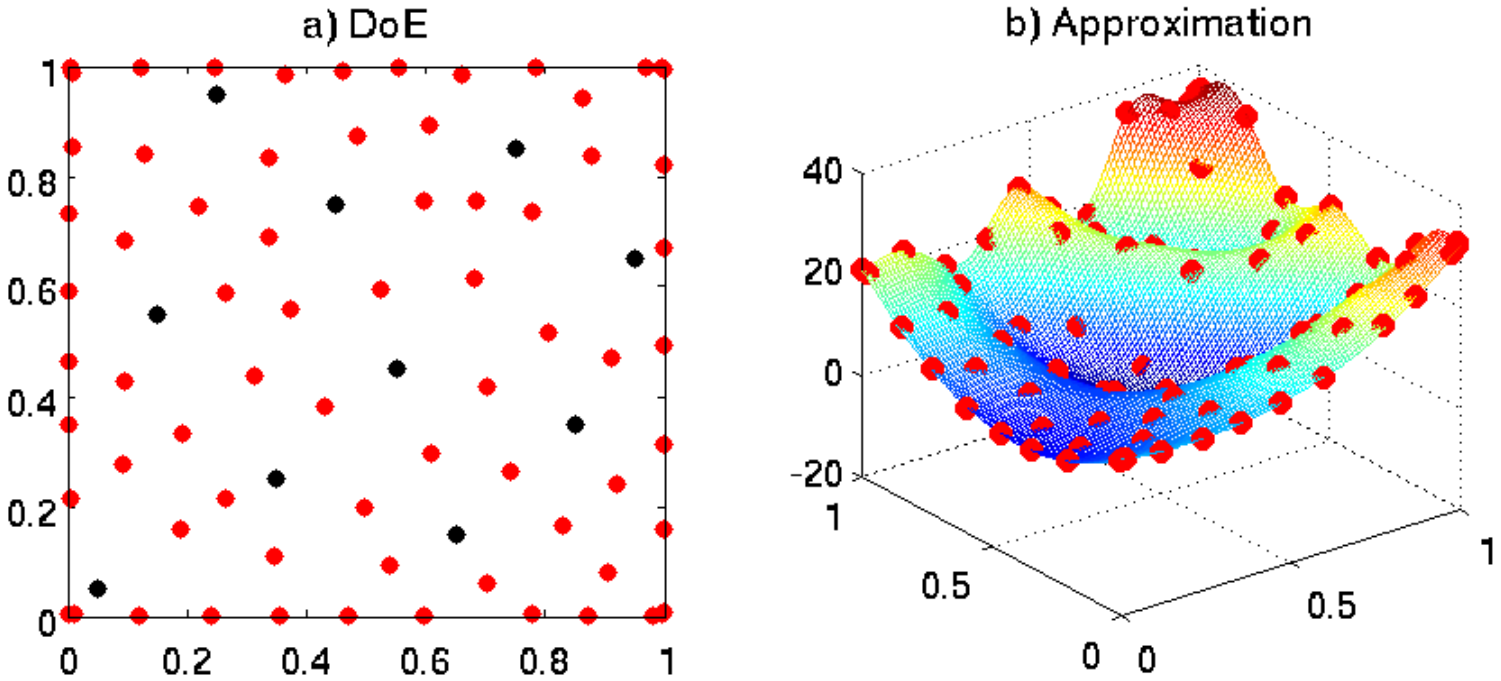


Рис.: Финальная обучающая выборка и аппроксимация

Пример: результаты

Адаптивный процесс DoE даёт очевидное снижение ошибки, и итоговая модель очень точная.

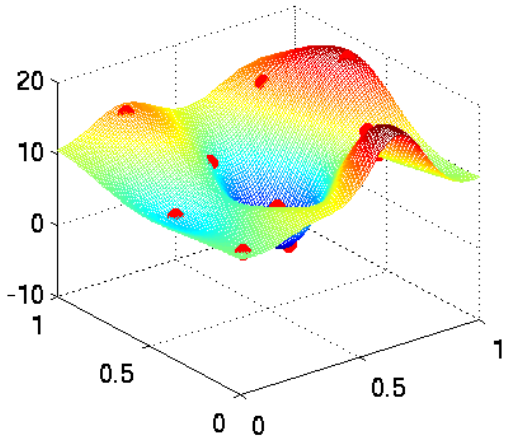


Рис.: Изначальная модель

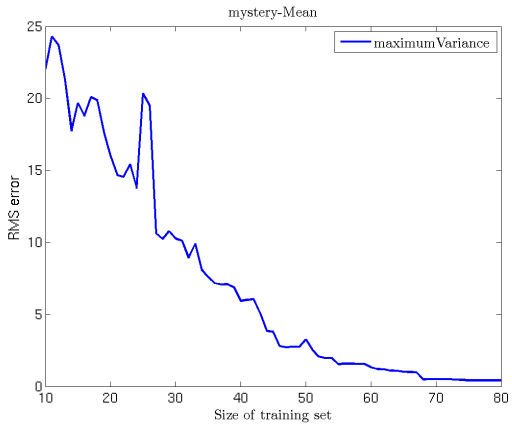


Рис.: Зависимость ошибки от размера обучающей выборки

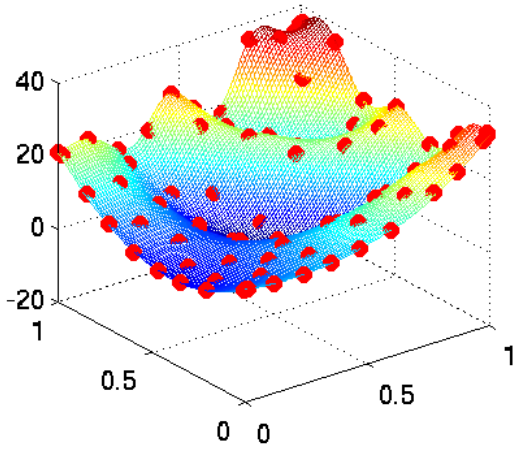


Рис.: Финальная модель