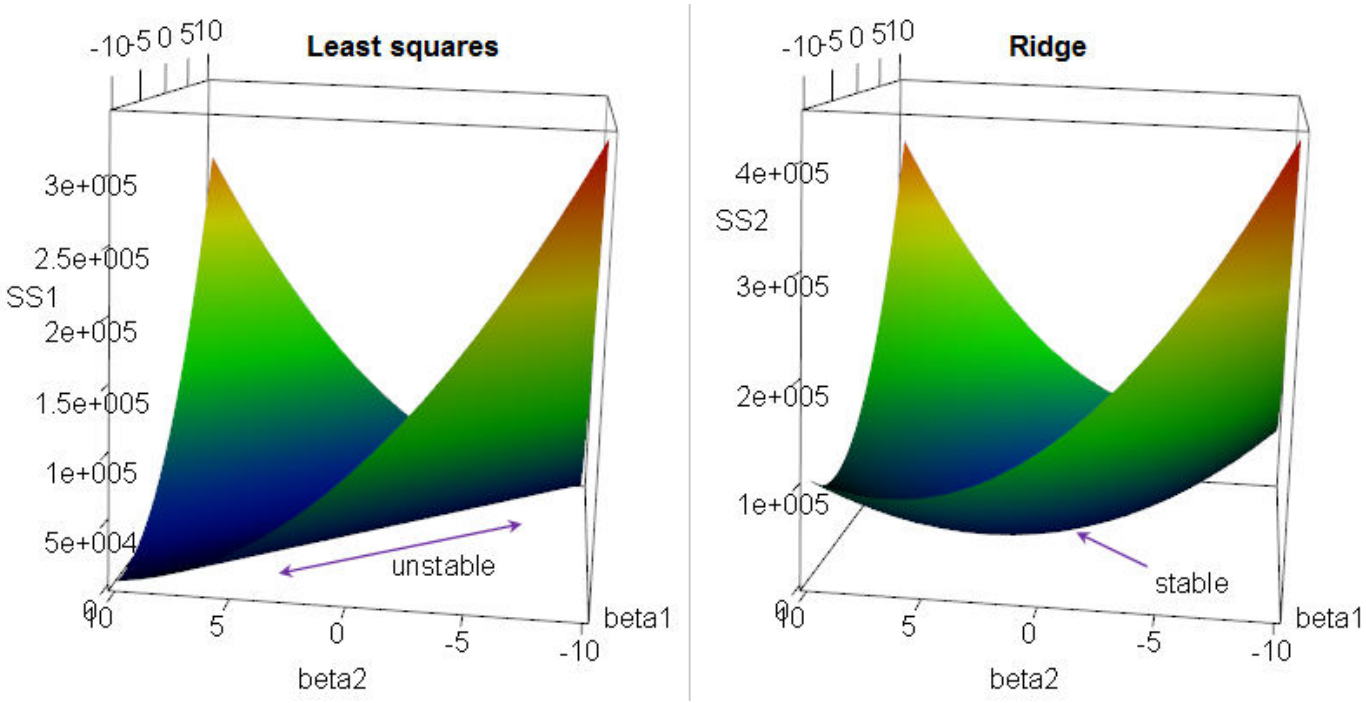
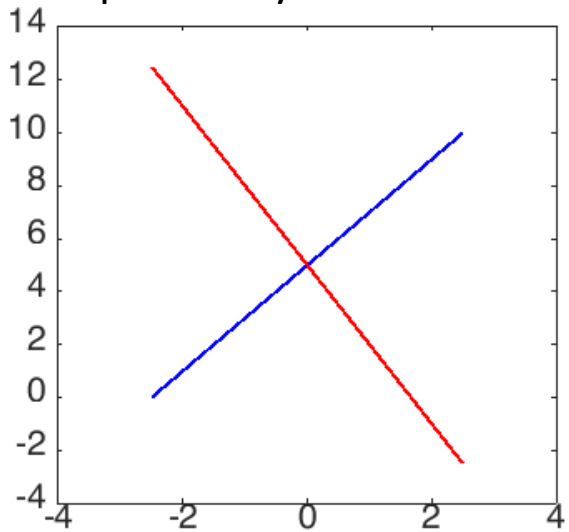
Занятие 5. Мультиколлинеарность и регуляризация Тихонова



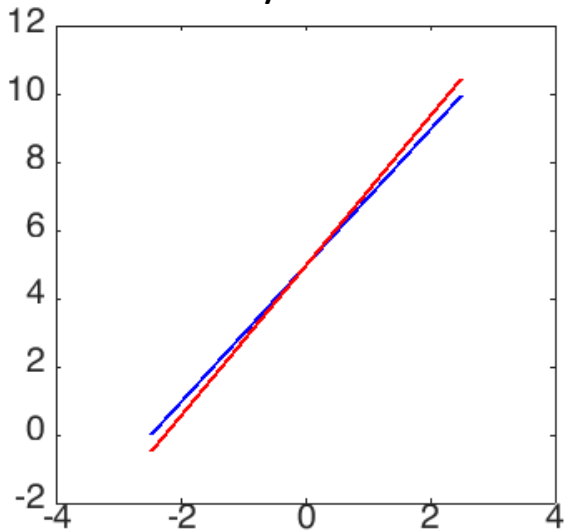
- Обусловленность системы линейных уравнений; плохо обусловленные системы
- Метод регуляризации Тихонова
- Понятие о мультиколлинеарности
- Гребневая регрессия и выбор параметра регуляризации

Обусловленность системы линейных уравнений

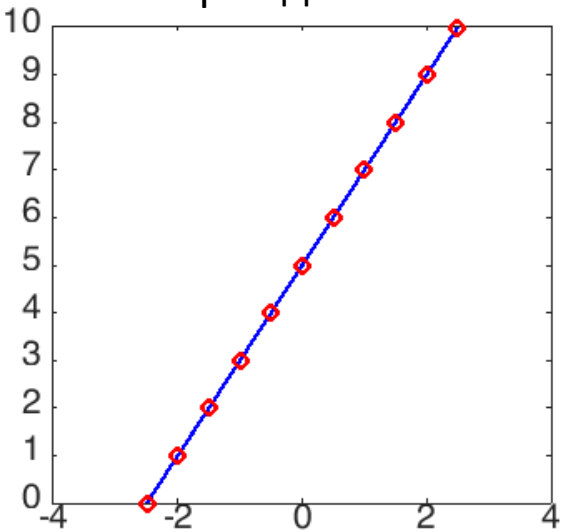
Хорошо обусловленная



Плохо обусловленная



Вырожденная



Проблемы, связанные с плохо обусловленными задачами

- Чувствительность решения к ошибкам в исходных данных
- Возможно сильное влияние погрешностей численных расчётов

Возможное решение проблемы – регуляризация (привнесение в задачу дополнительной информации о свойствах решения)

Метод регуляризации Тихонова

Исходная система уравнений

$$Ax = y$$

$\det A \approx 0$ (плохо обусловленная)

Регуляризованная система уравнений

$$(A + \lambda I)x = y$$

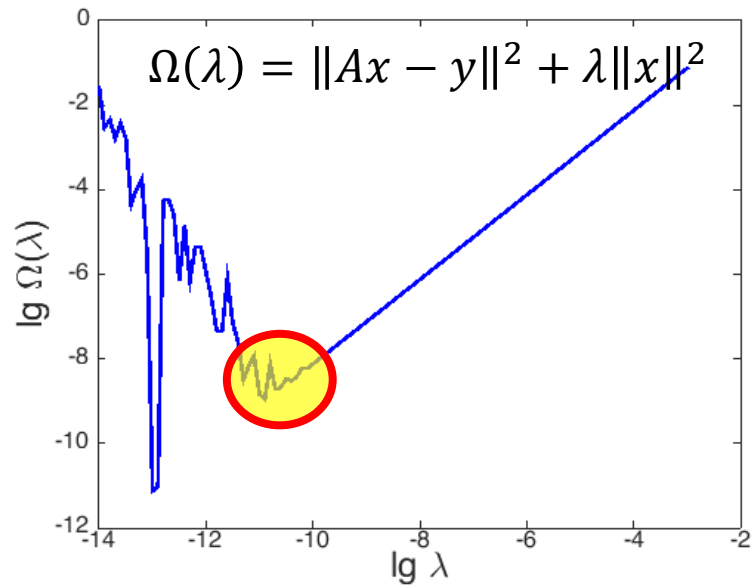
$\lambda > 0$ – параметр регуляризации

Пример плохо обусловленной системы линейных уравнений и проблем с ее решением

$$\begin{cases} x + 7y = 5 \\ \sqrt{2}x + \sqrt{98}y = \sqrt{50} \end{cases}$$

```
>> A = [1 7; 2^0.5 98^0.5];
>> y = [5; 50^0.5];
>> x = inv(A)*y; disp(x');
warning: matrix singular to
machine precision
    0.60938    1.24121
>> disp((A*x)')
    9.2979    13.1491
```

Поиск оптимального λ (минимизация функционала Тихонова)



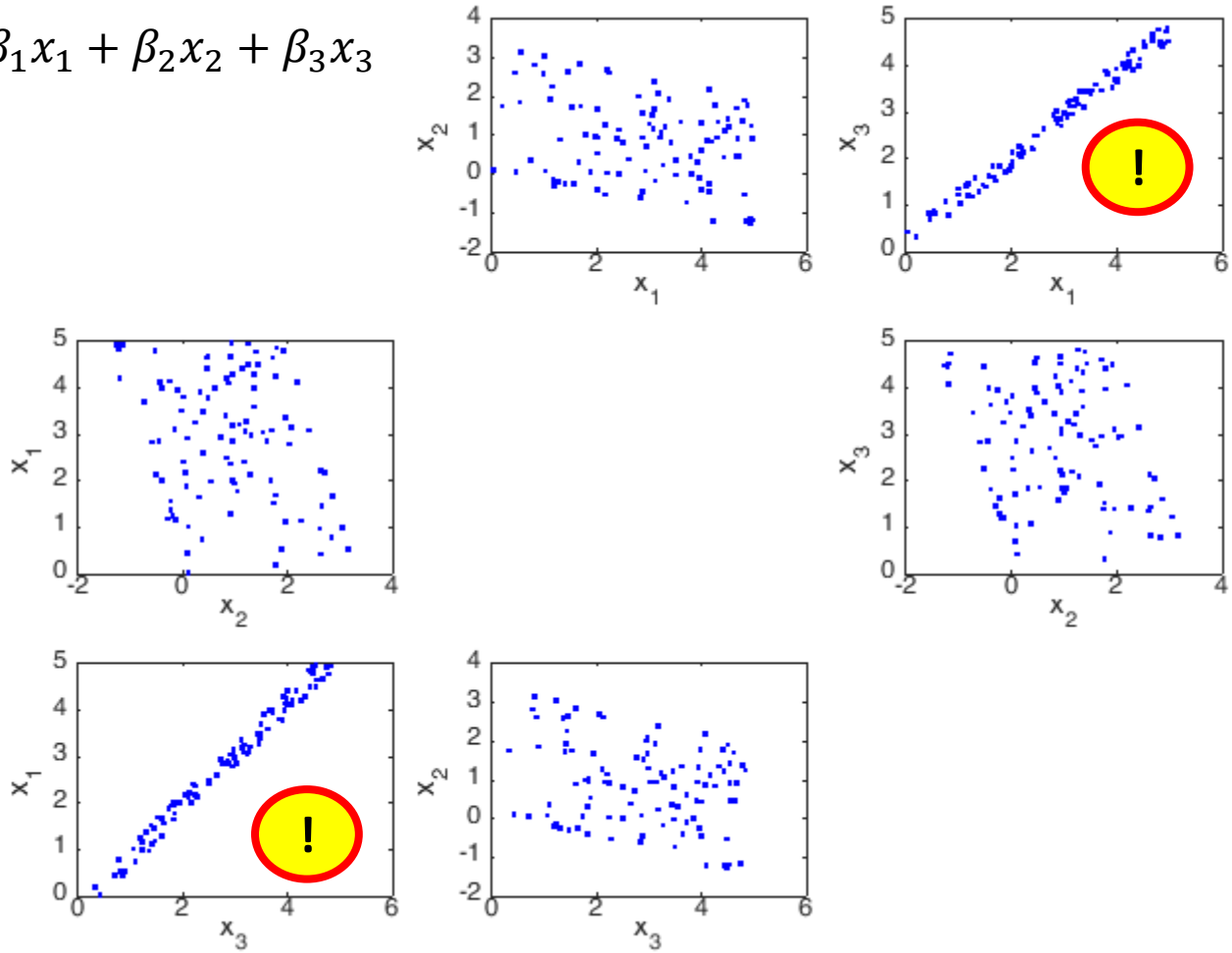

```
>> k = 1e-9;
>> x=inv(A+k*eye(2))*y;disp(x);
    0.45874    0.64875
>> disp((A*x)')
    5.0000     7.0711
```

Мультиколлинеарность

Мультиколлинеарность – наличие линейной зависимости между объясняющими переменными (факторами) регрессионной модели

Графики зависимости факторов друг от друга

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$



Гребневая регрессия (Ridge regression)

Метод наименьших квадратов

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

Минимизация функции

$$RSS = \sum_i (y_i - \hat{y}_i)^2$$

Гребневая регрессия

$$\tilde{\beta} = (X^\top X + \lambda I)^{-1} X^\top y$$

Минимизация функции

$$RSS = \sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_j \beta_j^2$$

Смещенность оценки $\tilde{\boldsymbol{\beta}}$

$$E[\tilde{\beta}] = E[(X^\top X + \lambda I)^{-1} X^\top (XB + \varepsilon)] = (X^\top X + \lambda I)^{-1} (X^\top X) B$$

Дисперсия оценки $\tilde{\boldsymbol{\beta}}$

$$\begin{aligned}\operatorname{cov}(\tilde{\beta}, \tilde{\beta}) &= E\left[(X^\top X + \lambda I)^{-1} X^\top \varepsilon \left((X^\top X + \lambda I)^{-1} X^\top \varepsilon\right)^\top\right] \\ &= (X^\top X + \lambda I)^{-1} X^\top E[\varepsilon \varepsilon^\top] X (X^\top X + \lambda I)^{-1} \\ &= \sigma^2 (X^\top X + \lambda I)^{-1} X^\top X (X^\top X + \lambda I)^{-1}\end{aligned}$$

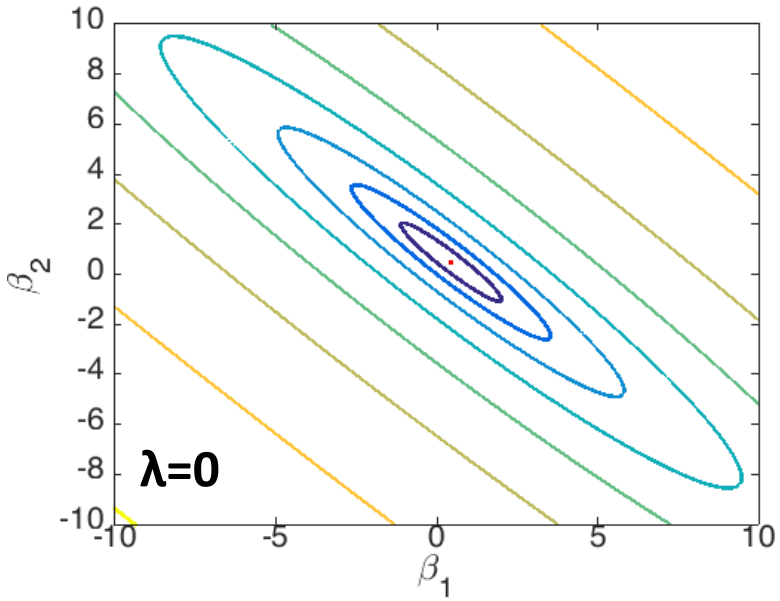
$\sigma^2 = \sqrt{(e^\top e)/(n-k)}$ - ошибка регрессии

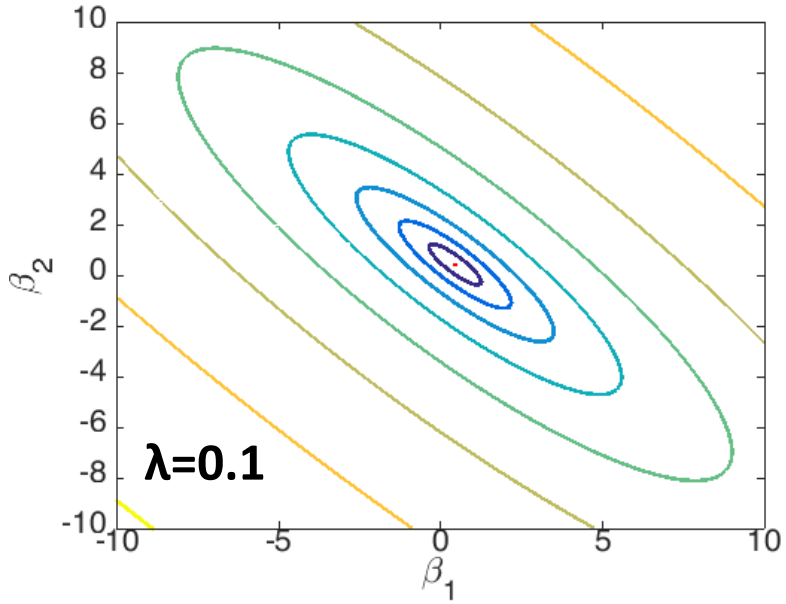
Дисперсии $\tilde{\beta}$ - на главной диагонали матрицы

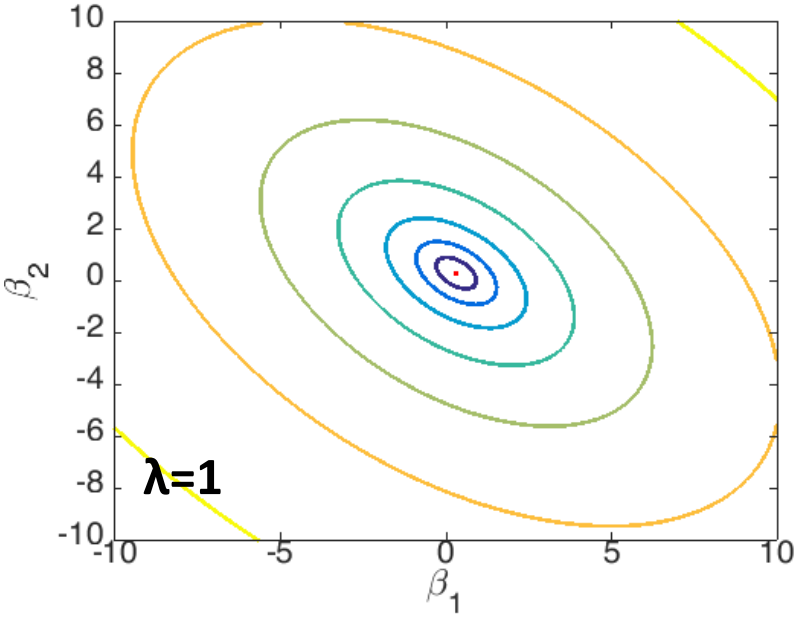
Гребневая регрессия: поверхность функции RSS(β)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Обратите внимание: при росте параметра регуляризации эллипсы изолиний вокруг минимума «стягиваются» в окружности







Гребневая регрессия: стандартизация точек

Исходная модель

$$y_i = \beta_0 + \sum_j \beta_j x_{ij}$$

$$RSS = \sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_j \beta_j^2$$

НО!

- Разная размерность у разных β_i и x_j
- При $\lambda \to \infty$ $\beta_0 \to 0$ (нежелательно)

Необходимость нормировки

Стандартизация (нормировка) исходных данных

$$y_i^* = \frac{y_i - \bar{y}}{\sqrt{\sum_j (y_j - \bar{y})^2}} = \frac{y_i - \bar{y}}{y_{norm}}$$

$$x_{ij}^* = \frac{x_{ij} - \overline{x_j}}{\sqrt{\sum_k (x_{kj} - \overline{x_j})^2}} = \frac{x_{ij} - \overline{x_j}}{x_{j,norm}}$$

Переход к исходным размерностям

$$y_i^* = \beta_0^* + \sum_j \beta_j^* x_{ij} \Rightarrow \frac{y_i - \bar{y}}{y_{norm}} = \beta_0^* + \sum_j \beta_j^* \frac{x_{ij} - \overline{x_j}}{x_{j,norm}} \Rightarrow$$

=0

$$y_i = \left(y_{norm}\beta_0^* + \bar{y} - y_{norm} \sum_j \frac{\beta_j^* \overline{x_j}}{x_{j,norm}} \right) + y_{norm} \sum_j \frac{\beta_j^* x_{ij}}{x_{j,norm}}$$

Гребневая регрессия: фактор инфляции дисперсии

VIF (Variance Inflation Factor, фактор инфляции дисперсии) – мера мультиколлинеарности, позволяет оценить увеличение дисперсии из-за линейной зависимости фактора x_i от остальных

$$\mathrm{Var}(\beta_j) = \frac{\sigma^2}{\sum_i (x_{ij} - \overline{x_j})^2} VIF_j = \frac{\sigma^2}{\sum_i (x_{ij} - \overline{x_j})^2} \frac{1}{1 - R_j^2}$$

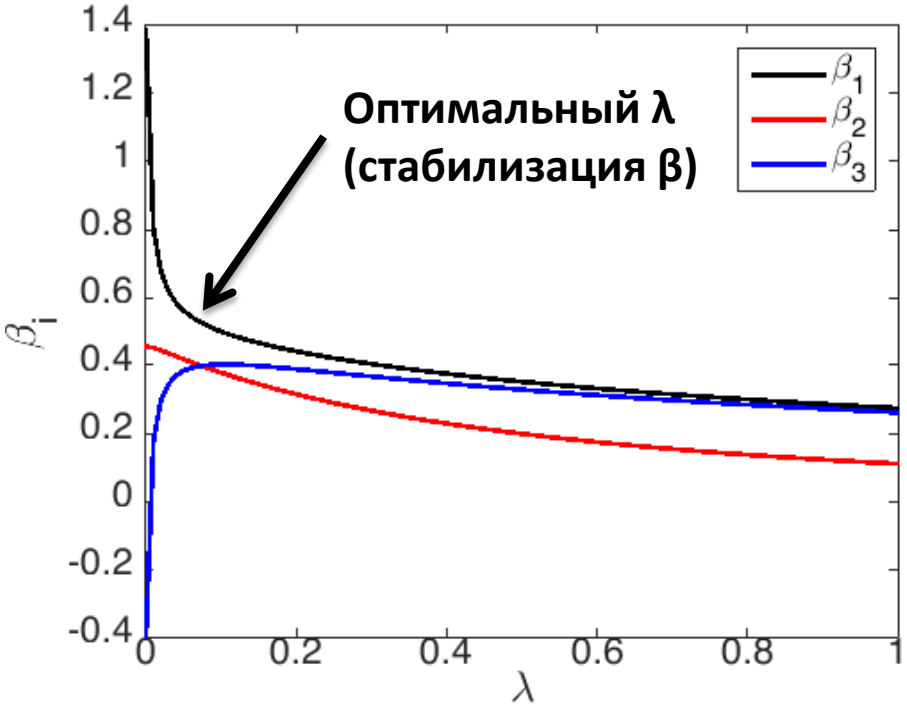
R_j^2 - коэффициент детерминации линейной зависимости фактора x_i от остальных факторов

Возможные значения VIF

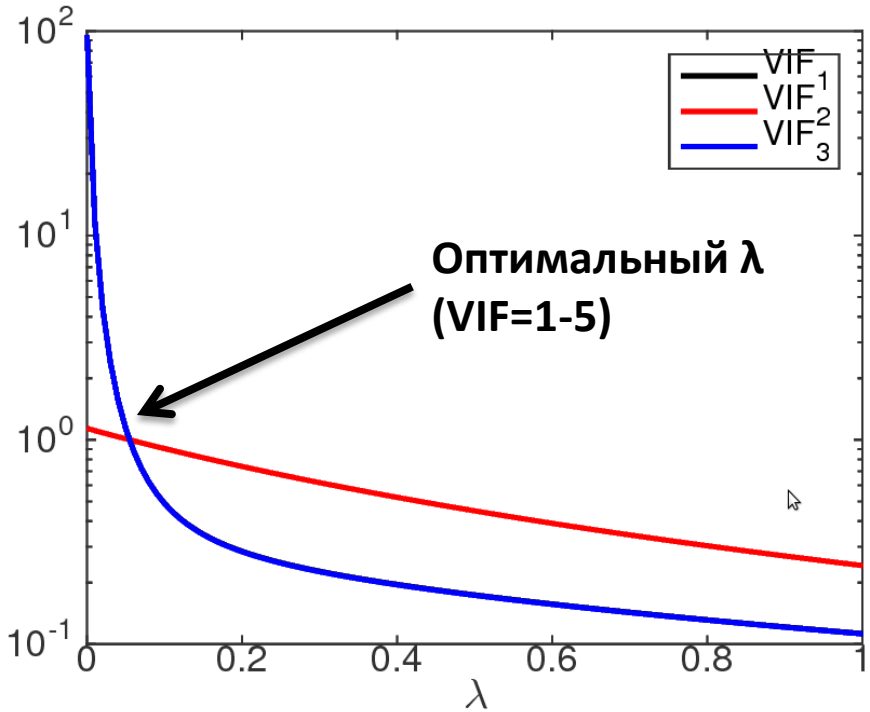
- VIF > 10 – выраженная мультиколлинеарность
- VIF = 5-10 – мультиколлинеарность
- VIF = 1-5 – нет мультиколлинеарности

Гребневая регрессия: подбор параметра регуляризации

Ridge trace plot
(график следа гребня)



VIF plot
(фактор инфляции дисперсии)



$$\mathrm{Var}(\beta_j) = \frac{\sigma^2}{\sum_i (x_{ij} - \overline{x_j})^2} VIF_j$$