My first scientific paper

Week 8

Bayesian inference

Vadim Strijov

Moscow Institute of Physics and Technology

2021

Гипотеза порождения данных для линейной модели

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

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

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

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

Функция правдоподобия данных

Функция вероятности появления зависимой переменной имеет вид

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

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

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

Коэффициент Z_D определен выражением, нормирующим функцию плотности нормального распределения

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

Функция правдоподобия данных при $B = \beta I$

Для гомоскедаксичного случая функция ошибки равна

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

а нормирующий множитель

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

Априорное (sic!) распределение параметров модели

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

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

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

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

Нормирующая константа $Z_{\mathbf{w}}$ равна

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

При равенстве дисперсий элементов вектора параметров

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

Байесовский вывод, первый уровень

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

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

Элементы этого выражения и соответствующие им параметры:

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

Апостериорное распределение параметров, частный случай

Апостериорное распределение параметров модели для заданных матриц A, B

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

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

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

получаем вместо вышестоящего выражение

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

где Z_S — нормирующий множитель.

Апостериорное распределение параметров, частный случай

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

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

а функция ошибки —

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

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

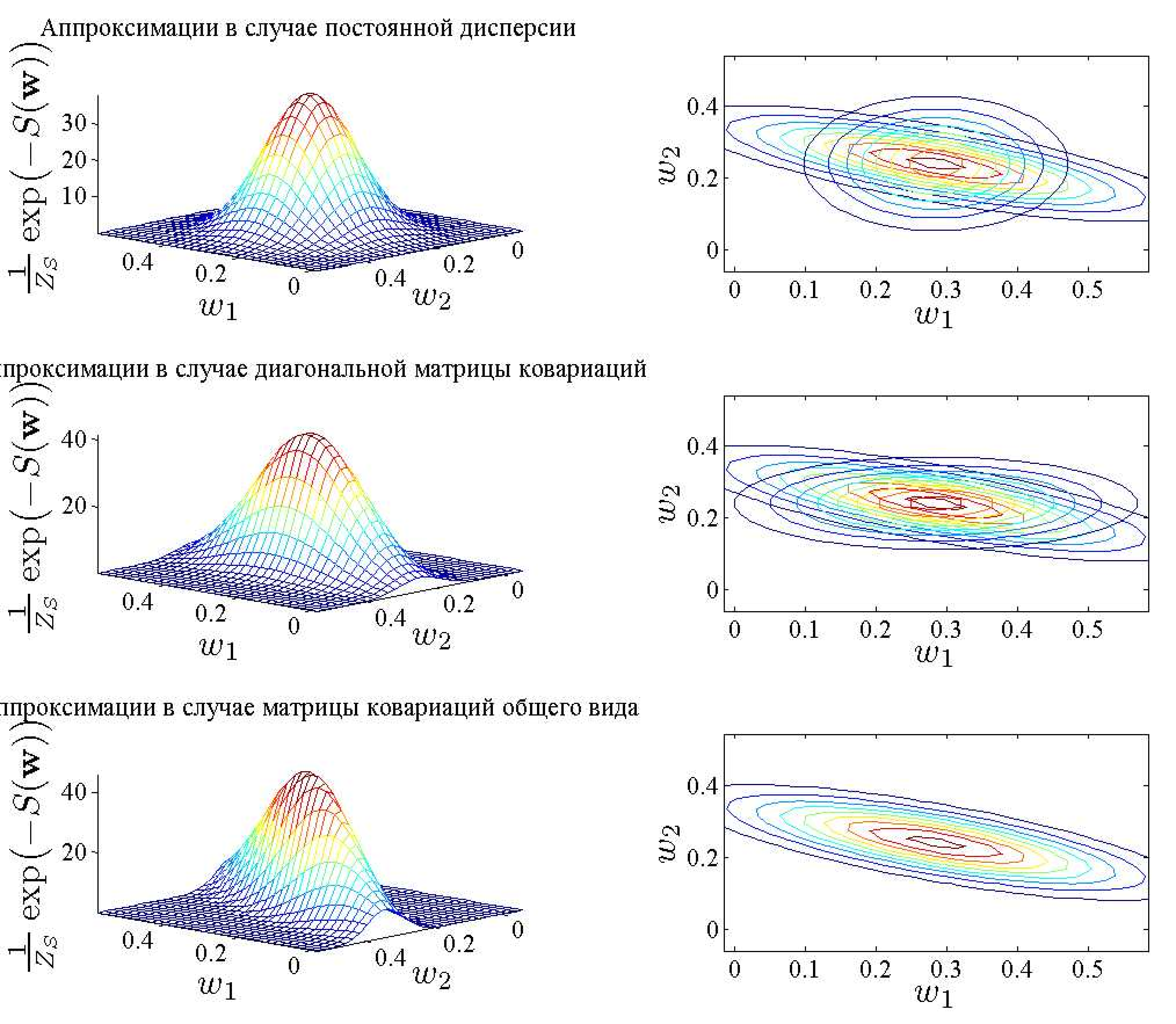
Функция ошибки включает две матрицы ковариации

Согласно первому уровню Байесовского вывода

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

Имеется девять возможных вариантов гипотезы порождения данных.

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


Аппроксимации в случае постоянной дисперсии
ппроксимации в случае диагональной матрицы ковариаций
ппроксимации в случае матрицы ковариаций общего вида
$\frac{1}{Z_S}\exp(-S(\mathbf{w}))$
w_1
w_2

Наиболее вероятные параметры и правдоподобие модели

Наиболее вероятные параметры

$$\mathbf{w}_{\text{MP}} = \arg\max_{\mathbf{w}\in\mathcal{W}} p(\mathbf{w}|D, f, A, B),$$

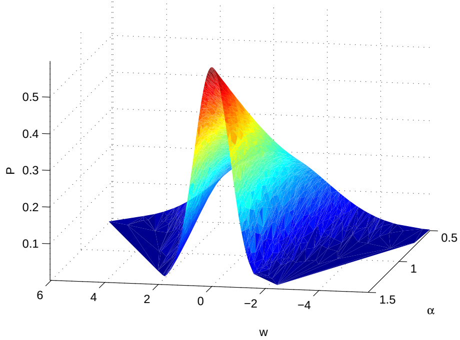
модели f оцениваются посредством Байесовского подхода:

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

причем функция правдоподобия данных $p(D|\mathbf{w}, f, B)$ определена гипотезой порождения зависимой переменной $\mathbf{y}$. Правдоподобие модели (evidence):

$$\mathcal{E}\left(f(\mathbf{w}, \mathbf{x})\right) = \int p(D|\mathbf{w}, f, B)p(\mathbf{w}|f, A)d\mathbf{w}.$$

Зависимость распределения параметров от $A = \alpha I_n$



- z-axis: $p(\mathbf{w}|D, f, A, B)$ — распределение параметров,
- y-axis: α — обратное значение дисперсии,
- x-axis: w — параметр модели.

Многоуровневые модели и индексация объектов и признаков

Заданы индексы

- объектов $\{1,\dots,i,\dots,m\} = \mathcal{I}$, разбиение$\mathcal{I} = \mathcal{B}_1 \sqcup \dots \sqcup \mathcal{B}_K$;
- признаков $\{1,\dots,j,\dots,n\} = \mathcal{J}$, активный набор $\mathcal{A} \subseteq \mathcal{J}$.

Регрессионная модель

$$f : (\mathbf{w}, \mathbf{x}) \mapsto y;$$

Выбираемая модель задана набором индексов $\mathcal{A}$:

$$\mathbb{E}(\mathbf{y}|X) = X_{\mathcal{A}}\mathbf{w}_{\mathcal{A}}, \text{ or } \mathbb{E}(y_i|\mathbf{x}) = \mathbf{w}_{\mathcal{A}}^{\mathsf{T}}\mathbf{x}_i.$$

Многоуровневая модель $\mathfrak{f}$ — это набор моделей $\mathfrak{f} = \{f_k | k = 1,\dots,K\}$, такой, что для каждого значения k

$$\mathbb{E}(y_{i\in\mathcal{B}_k}|\mathbf{x}) = \mathbf{w}_{(k)}^{\mathsf{T}}\mathbf{x}_{i\in\mathcal{B}_k},$$

при разбиении

$$\mathcal{I} = \sqcup_{k=1}^{K}\mathcal{B}_k \ni i.$$

Задача выбора одной модели или их набора

Выбор модели:

$$\hat{f}(\mathbf{w}, \mathbf{x}) = \arg\max_{\mathcal{A} \subseteq \mathcal{J}} \mathcal{E}\left(f(\mathbf{w}_{\mathcal{A}}, \mathbf{x})\right).$$

Выбор многоуровневой модели:

$$\hat{\mathfrak{f}}(\mathbf{w}_{(1)}, \dots, \mathbf{w}_{(K)}, \mathbf{x}) = \arg\max_{\sqcup_{k=1}^{K} \mathcal{B}_k = \mathcal{I}} \prod_{k=1}^{K} \mathcal{E}\left(f(\mathbf{w}_{(k)}, \mathbf{x}_{\mathcal{B}_k})\right).$$

William of Ockham, 1288–1348 (University of Oxford, 1309–1321)

Entia non sunt multiplicanda praeter necessitatem.

Бритва Оккама: не умножай сущности без необходимости.

Связанный Байесовский вывод — метод выбора моделей

Метод использует Байесовский вывод дважды:

1. при оценке апостериорного распределения параметров моделей и
2. при оценке апостериорной вероятности модлей.

Байесовское сравнение моделей, второй уровень вывода

Рассмотрим набор конкурирующих моделей $f_1, \ldots, f_K$, приближающих данные D. Обозначим $p(f_k)$ априорную вероятность k-й модели. Апостериорная вероятность

$$p(f_k|D) = \frac{p(D|f_k)p(f_k)}{\sum_{q=1}^{K} p(D|f_q)p(f_q)}.$$

Функция $p(D|f_k)$ от данных D, при фиксированной модели f_k называется правдоподобием (evidence) этой модели. Так как знаменатель $p(D) = \sum\limits_{q=1}^{K} p(D|f_q)p(f_q)$ не зависит от модели, то модели сравниваются посредством правдоподобия

$$\frac{p(f_k|D)}{p(f_q|D)} = \frac{p(f_k)p(D|f_k)}{p(f_q)p(D|f_q)},$$

при допущении равенства их априорной вероятности $p(f_k) = p(f_q)$.

Пример вычисления правдоподобия модели

Let there be given the series $\{-1, 3, 7, 11\}$. One must to forecast the next two elements.
The model f_a:

$$x_{i+1} = x_i + 4$$

gives the next elements $15, 19$.
The model f_c:

$$x_{i+1} = -\frac{x_i^3}{11} + \frac{9x_i^2}{11} + \frac{23}{11}$$

gives the next elements $-19.9, 1043.8$.

Let the prior probabilities be equal or comparable.
Let each parameter of the models is in the set

$$\{-50, \dots, 0, \dots, 50\}.$$

A toy example, continued

The parameters $(n = 4, x_1 = -1)$ brings the proper model with zero-error.
The evidence of the model f_a is

$$p(D|f_a) = \frac{1}{101}\frac{1}{101} = 0.00010.$$

Let the denominators of the second models are in the set $\{0, \dots, 50\}$.
Take account of $c = -1/11 = -2/22 = -3/33 = -4/44$.
The evidence of the model f_c is

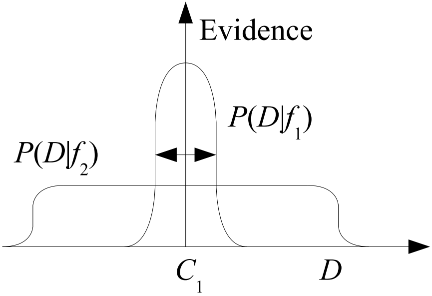
$$p(D|f_c) = \left(\frac{1}{101}\right)\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{4}{101}\frac{1}{50}\right) = 4.9202\ldots\times 10^{-12}.$$

The result of the model comparison is

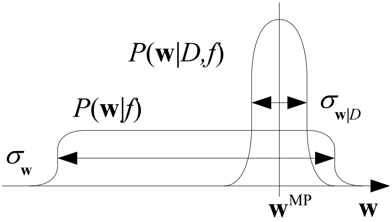
$$\frac{p(D|f_a)}{p(D|f_c)} = \frac{0.00010}{2.5\times 10^{-12}}.$$

Бритва Оккама

Если If f_2 — is more complex model, then its distribution $p(D|f_2)$ has smaller values (variance has greater values). If the errors of both models are equal, then the simple model f_1 is more probable than the complex model f_2.



Множитель Оккама



Множитель Оккама задан отношением дисперсий параметров модели.
Дисперсия $\sigma_{w|D}$ зависит от апостериорного распределения параметров $\mathbf{w}$.
Множитель Оккама отражает «сжатие» пространства параметров при появлении данных.

Как сравнить три модели, пример