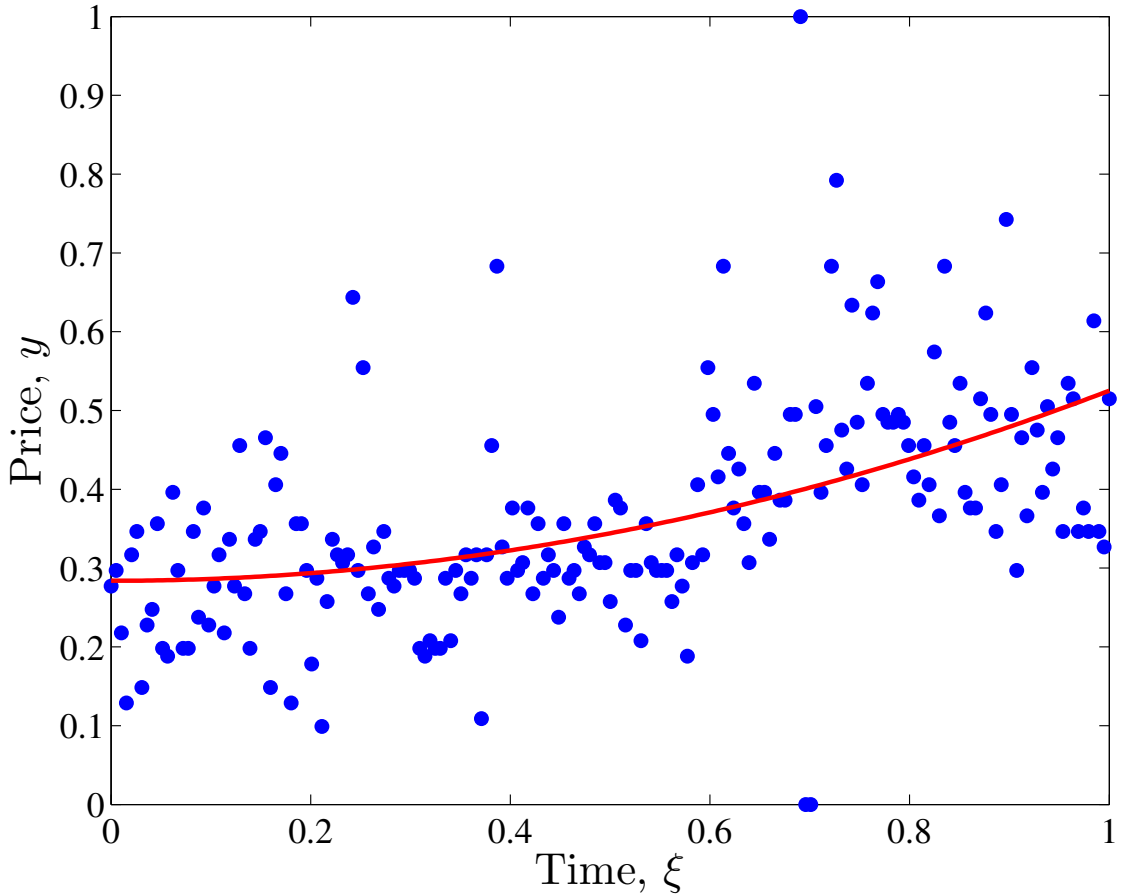
Mathematical methods of prediction: *Introduction to the probabilistic models*

Vadim Strijov

Moscow Institute of Physics and Technology

September 3, 2020

Regression model and regression function



The regression model is $f = \mathbf{x}^{\mathsf{T}}\mathbf{w}$, where $x_1 = \xi^0, x_2 = \xi^2$.

The regression function: $y = w_1 + w_2\xi^2 + \varepsilon(\xi)$

with the optimal parameters $\mathbf{w}_0 = [0.2839, 0.2412]^{\mathsf{T}}$.

The neural network: $f = \sigma'\left(\mathbf{w}^{\mathsf{T}}\boldsymbol{\sigma}\left(\mathbf{W}\mathbf{x} + \mathbf{b}\right) + b'\right)$.

Probabilistic model

Call the (approximation) model f the parametric family

$$f = f(\mathbf{w}, \mathbf{x}).$$

Call the residue $\varepsilon = f - y$. Assume, for example,

$$\varepsilon \sim \mathcal{N}\left(f, \frac{1}{\beta}\right), \quad \mathbf{w} \sim \mathcal{N}\left(\hat{\mathbf{w}}, \frac{1}{\alpha}\right).$$

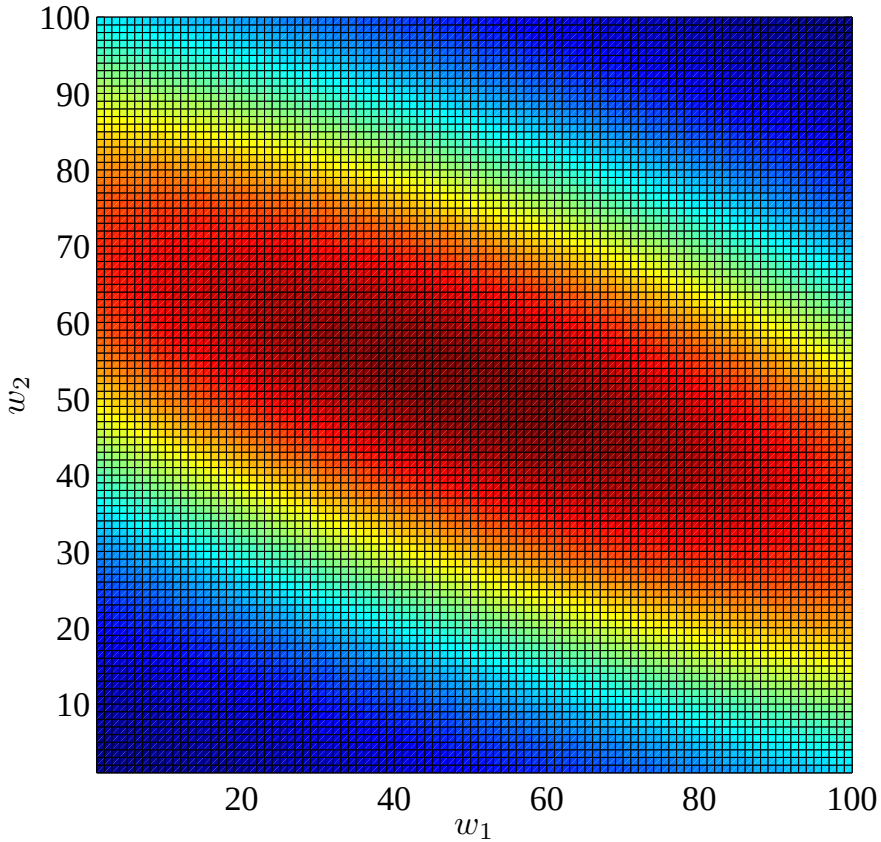
Call the probabilistic model the distribution

$$p(f|\mathbf{x}, \mathbf{w}).$$

The forecast is the expected value (at some point $\mathbf{x}_0$)

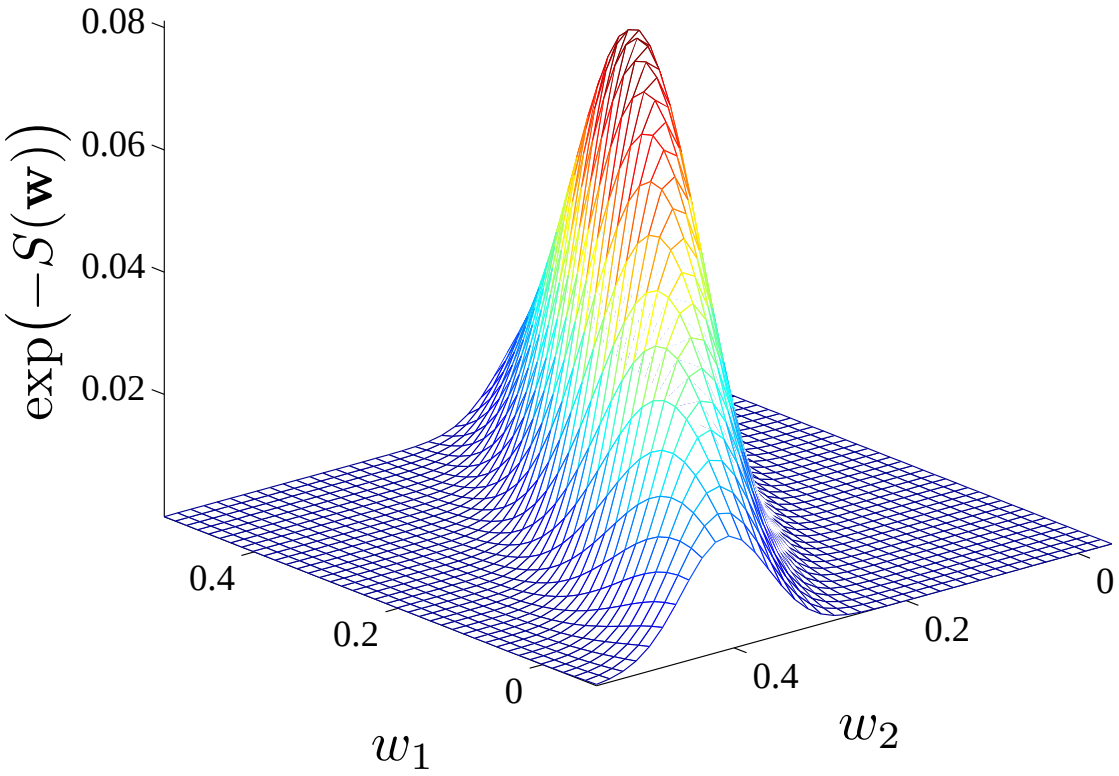
$$\mathrm{E}(f|\mathbf{x}_0, \hat{\mathbf{w}}).$$

Пространство параметров модели



Сэмплирование параметров модели полным перебором. Цветом обозначено значение функциии $\exp\big(-S(\mathbf{w})\big)$.

Пространство параметров модели



Вид функции $\exp\big(-S(\mathbf{w})\big)$ в окрестности вектора оптимальных параметров w_0.

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,
$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \mathrm{diag}(\beta_1, \ldots, \beta_m)^{-1} I)$$
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

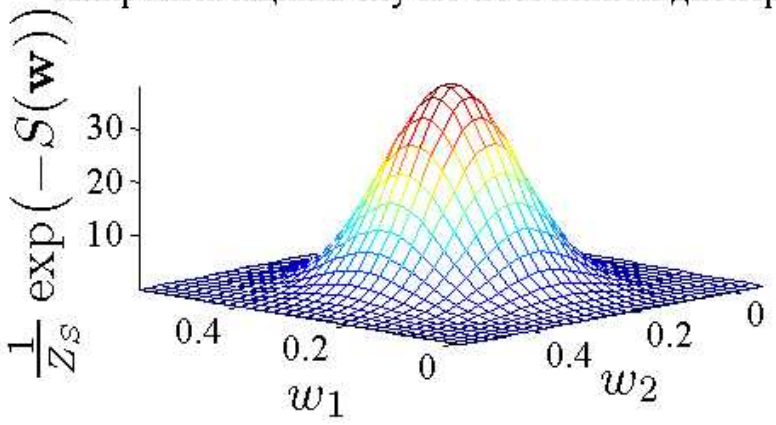
Согласно первому уровню Байесовского вывода

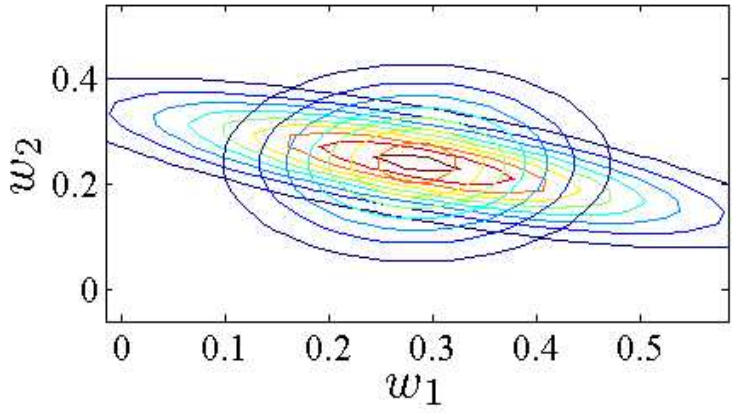
$$S(\mathbf{w}|D, f) = \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MP}})^{\mathsf{T}} A(\mathbf{w} - \mathbf{w}_{\text{MP}}) + \frac{1}{2}(\mathbf{f} - \mathbf{y})^{\mathsf{T}} B(\mathbf{f} - \mathbf{y}).$$

Имеется девять возможных вариантов гипотезы порождения данных.

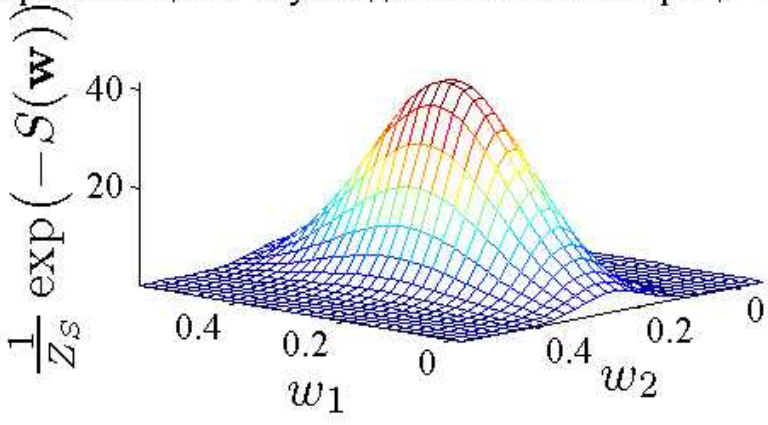
Обратная ковариационная матрица параметров	Обратная ковариационная матрица зависимой переменной
$A = \alpha I_n$	$B = \beta I_m$
$A = \text{diag}(\alpha_1, \ldots, \alpha_n)$	$B = \text{diag}(\beta_1, \ldots, \beta_m)$
A	B

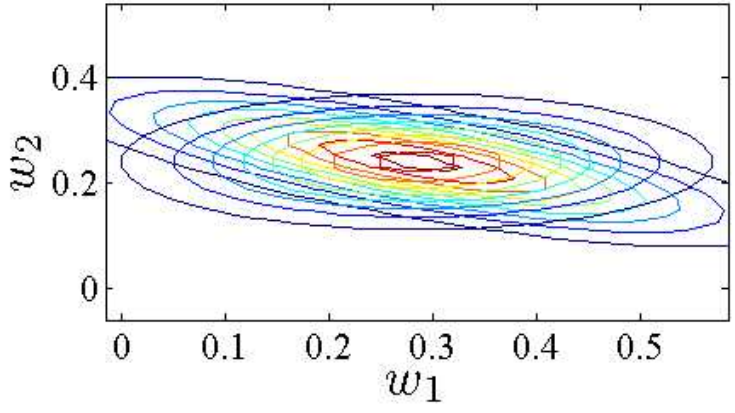
Аппроксимации в случае постоянной дисперсии



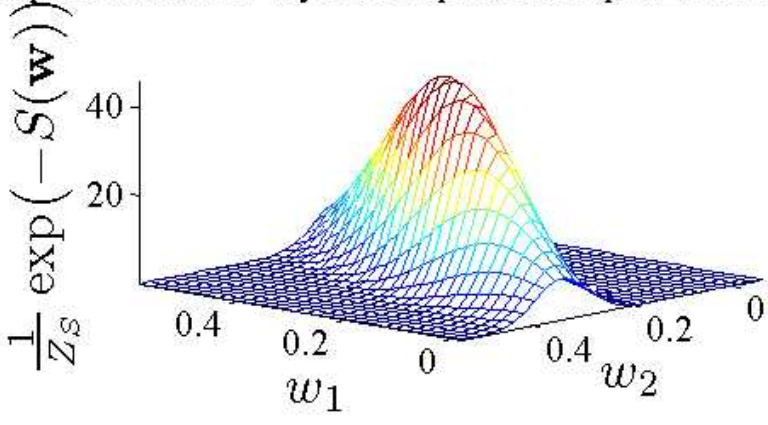


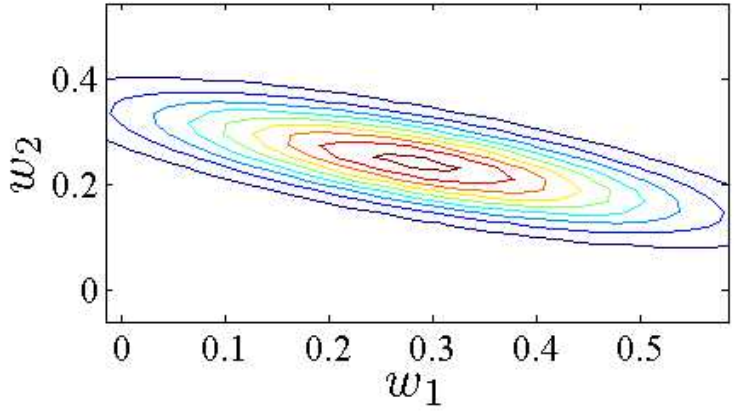
проксимации в случае диагональной матрицы ковариаций





проксимации в случае матрицы ковариаций общего вида





Аппроксимация Лапласа, одномерное распределение

Пусть задано ненормированное распределение $p^*(w)$.
Требуется найти нормировочную константу

$$Z_w = \int_{-\infty}^{\infty} p(w)^* dw,$$

при которой распределение $p(w) = Z_w^{-1} p^*(w)$. Предполагается, что $p^*(w)$ имеет максимум (моду м.с.в. $\mathbf{w}$) в точке w_0:

$$\left.\frac{dp(\mathbf{w})}{dw}\right|_{w=w_0} = 0.$$

Логарифмируя и разлагая $p^*(w)$ в ряд Тейлора в окрестности w_0, получим

$$\ln p^*(w) = \ln p^*(w_0) + 0 - \frac{\alpha}{2}(w - w_0)^2 + \cdots,$$

$$\text{где} \quad \alpha = -\left.\frac{\partial^2 \ln p^*(w)}{\partial w^2}\right|_{w=w_0}.$$

Аппроксимация Лапласа, одномерное распределение

Беря экспоненту обеих частей разложения получим,

$$p^*(w) \approx p^*(w_0) \exp\left(-\frac{\alpha}{2}(w - w_0)^2\right).$$

Тогда нормальное распределение $\hat{p}(w)$, приближающее нормированное распределение $p(w)$ имеет вид

$$\hat{p}(w) = \frac{1}{\sqrt{2\pi\alpha^{-1}}} \exp\left(-\frac{\alpha}{2}(w - w_0)^2\right),$$

а нормировочная константа для $p^*(w)$ —

$$Z_w \approx p^*(w_0)\sqrt{\frac{2\pi}{\alpha}}.$$

Аппроксимация Лапласа, многомерное распределение

Для отыскания $p(\mathbf{w}) = Z_{\mathbf{w}}^{-1} p^*(\mathbf{w})$ разложим в ряд Тейлора в окрестности $\mathbf{w}_0$ логарифм

$$\ln p^*(\mathbf{w}) = \ln p^*(\mathbf{w}_0) + 0 - \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)A(\mathbf{w} - \mathbf{w}_0) + \cdots ,$$

где матрица Гессе $A = [\alpha_{ij}]$ определена как

$$\alpha_{ij} = - \left. \frac{\partial^2 \ln p^*(\mathbf{w})}{\partial w_i \partial w_j} \right|_{\mathbf{w}=\mathbf{w}_0},$$

кратко,

$$A = -\nabla^2 \ln p^*(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_0}, \quad \text{здесь } \nabla \text{ — градиент функции.}$$

Аппроксимация Лапласа, многомерное распределение

Беря экспоненту разложения получим

$$p^*(\mathbf{w}) \approx p^*(\mathbf{w}_0) \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^\mathsf{T} A(\mathbf{w}-\mathbf{w}_0)\right).$$

Тогда нормальное распределение $\hat{p}(w)$, приближающее нормированное распределение $p(\mathbf{w})$ имеет вид

$$\hat{p}(w) = \mathcal{N}(\mathbf{w}_0, A^{-1}) = \frac{1}{(2\pi)^{\frac{n}{2}} \det^{-\frac{1}{2}} A} \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^\mathsf{T} A(\mathbf{w}-\mathbf{w}_0)\right),$$

а нормировочная константа для $p^*(w)$

$$Z_\mathbf{w} \approx p^*(\mathbf{w}_0) \frac{(2\pi)^{\frac{n}{2}}}{\det^{\frac{1}{2}} A}.$$