Заметки по матричным вычислениям и нормальному распределению

Матричные вычисления

Здесь и далее вектора будут обозначаться жирным шрифтом $\boldsymbol{x}, \boldsymbol{y}, \dots$, а матрицы – заглавными буквами $A, B, \dots$. При этом под вектором всегда будет пониматься вектор-столбец $\boldsymbol{x} = [x_1, \dots, x_d]^T$, индекс T обозначает транспонирование. Например, запись $\boldsymbol{x}^T A \boldsymbol{y}$ означает $\sum_{i,j} a_{ij} x_i y_j$.

Использование матрично-векторной нотации часто оказывается удобным с точки зрения вывода формул в многомерном пространстве. Кроме того, матричная запись позволяет в некоторых случаях повысить эффективность вычисления соответствующих формул на компьютере. Например, эффективность современных алгоритмов вычисления произведения двух матриц AB имеет сложность n^α, где n – размерность матрицы, а $\alpha < 2.42$.

Основные матричные операции представлены в таблице 1. Здесь под производной скалярной функции $f(\boldsymbol{x})$ по вектору $\boldsymbol{x}$ понимается градиент $\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = [\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \dots, \frac{\partial f(\boldsymbol{x})}{\partial x_d}]^T$, под производной скалярной функции $f(A)$ по матрице A – матрица частных производных $\left(\frac{\partial f(A)}{\partial A}\right)_{ij} = \frac{\partial f(A)}{\partial a_{ij}}$, под производной вектор-функции $\boldsymbol{f}(x)$ по скалярному аргументу x – вектор $\frac{\partial \boldsymbol{f}(x)}{\partial x} = [\frac{\partial f_1(x)}{\partial x}, \dots, \frac{\partial f_d(x)}{\partial x}]^T$, под производной матричной функции $A(x)$ по скалярному аргументу – матрица $\left(\frac{\partial A(x)}{\partial x}\right)_{ij} = \frac{\partial A_{ij}(x)}{\partial x}$.

Таблица 1: Основные матричные тождества

Базовые операции: $A(B+C) = AB + AC$, $(A+B)^T = A^T + B^T$, $(AB)^T = B^T A^T$, $(AB)^{-1} = B^{-1}A^{-1}$, $(A^{-1})^T = (A^T)^{-1}$.	След и определитель: $\det(AB) = \det A \det B$, $\det(A^{-1}) = 1/\det A$, $\det A = \prod_j \lambda_j$, $\operatorname{tr} A = \sum_j A_{jj} = \sum_j \lambda_j$, $\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB)$.
Производные следа и определителя: $\frac{\partial}{\partial A} \operatorname{tr} AB = B^T$, $\frac{\partial}{\partial A} \det A = (\det A)(A^{-1})^T$, $\frac{\partial}{\partial x} \log \det A(x) = \operatorname{tr}\left(A^{-1} \frac{\partial A}{\partial x}\right)$.	Производные: $\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{a} = \boldsymbol{a}$, $\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T A \boldsymbol{x} = (A + A^T)\boldsymbol{x}$, $\frac{\partial}{\partial A} \boldsymbol{x}^T A \boldsymbol{y} = \boldsymbol{x}\boldsymbol{y}^T$, $\frac{\partial}{\partial x} A^{-1} = -A^{-1} \frac{\partial A}{\partial x} A^{-1}$.

Используя свойства из таблицы 1, легко показать, например, что

$$\frac{\partial}{\partial \boldsymbol{x}} (\boldsymbol{t} - A\boldsymbol{x})^T (\boldsymbol{t} - A\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}} [\boldsymbol{t}^T \boldsymbol{t} - 2\boldsymbol{x}^T A^T \boldsymbol{t} + \boldsymbol{x}^T A^T A \boldsymbol{x}] = -2A^T \boldsymbol{t} + 2A^T A \boldsymbol{x} = 2A^T (A\boldsymbol{x} - \boldsymbol{t}).$$

Рассмотрим некоторые случаи использования матричных операций для повышения эффективности вычислений на компьютере. Пусть необходимо решить систему линейных уравнений с невырожденной квадратной положительно-определенной матрицей $A\boldsymbol{x} = \boldsymbol{t}$. Прямое решение данной системы $\boldsymbol{x} = A^{-1}\boldsymbol{t}$ требует вычисления обратной матрицы A^{-1}. Вычисление обратной матрицы – численно неустойчивая процедура, особенно для плохо обусловленных матриц. Однако, в данном случае нам не требуется знать полностью значение обратной матрицы, достаточно знать лишь вектор $A^{-1}\boldsymbol{t}$. Для его эффективного вычисления воспользуемся разложением Холецкого $A = RR^T$, где R – нижнетреугольная матрица, т.е. матрица, у которой все элементы выше главной диагонали равны нулю. Тогда

$$A\boldsymbol{x} = \boldsymbol{t} \;\Leftrightarrow\; R\underbrace{R^T\boldsymbol{x}}_{\boldsymbol{y}} = \boldsymbol{t} \;\Leftrightarrow\; \begin{cases} R\boldsymbol{y} = \boldsymbol{t}, \\ R^T\boldsymbol{x} = \boldsymbol{y}. \end{cases}$$

Таким образом, решение системы линейных уравнений $A\boldsymbol{x} = \boldsymbol{t}$ эквивалентно решению двух систем линейных уравнений с треугольными матрицами. Эти системы могут быть легко решены путем последовательного исключения неизвестных.

Другой пример связан с вычислением выражения $\log\det A$ для положительно-определенной матрицы A. Снова воспользуемся разложением Холецкого:

$$A = RR^T,\; \log\det A = \log\det RR^T = 2\log\det R = 2\sum_i \log R_{ii}.$$

Рассмотрим четыре матрицы $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{m\times m}$, $U \in \mathbb{R}^{n\times m}$, $V \in \mathbb{R}^{m\times n}$. Тогда справедливы следующие утверждения:

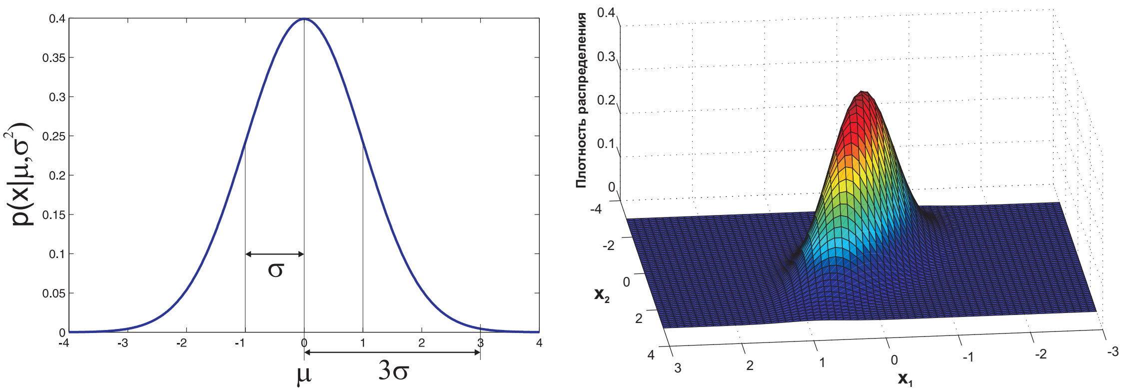
$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \text{ – тождество Вудбери,}$$
$$\det(A + UCV) = \det A \det C \det(C^{-1} + VA^{-1}U) \text{ – лемма об определителе матрицы.}$$

Предположим, что $n \gg m$ и величины $A^{-1}, \det A$ известны. Тогда данные утверждения позволяют свести вычисление обратной матрицы и определителя матрицы размера $n\times n$ к вычислению обратной матрицы и определителя матрицы размера $m\times m$.

Нормальное распределение

Случайная величина $x \in \mathbb{R}$ имеет нормальное (гауссовское) распределение с параметрами μ и σ^2, если ее плотность задается выражением (см. рис. 1,а)

$$p(x|\mu,\sigma^2) = \mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad \sigma > 0.$$



(а) (б)

Рис. 1: Плотность одномерного (а) и многомерного (б) нормальных распределений.

Параметры μ и σ^2 определяют, соответственно, мат.ожидание и дисперсию нормальной случайной величины.

По центральной предельной теореме среднее арифметическое независимых случайных величин с ограниченными мат.ожиданием и дисперсией стремится к нормальному распределению. Поэтому это распределение часто используется в качестве модели шума, который определяется суммой большого количества независимых друг от друга случайных факторов.

Из неравенства Чебышева известно, что для произвольной одномерной случайной величины с конечными мат.ожиданием и дисперсией вероятность отклонения от своего мат.ожидания на k

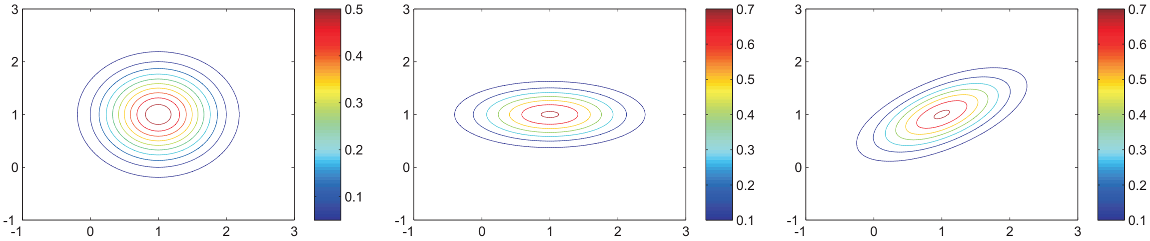


Рис. 2: Линии уровня различных нормальных распределений: (а) — нормальное распределение с матрицей ковариации, пропорциональной единичной $\Sigma = \alpha I$, (б) — нормальное распределение с диагональной матрицей ковариации, (в) — нормальное распределение с матрицей ковариации общего вида.

стандартных отклонений не превышает $1/k^2$. В частности, для $k = 3$ эта вероятность отклонения составляет 11.2%. Однако, для нормального распределения вероятность отклонения от своего мат.ожидания на 3 стандартных отклонения составляет всего 0.3%, что намного меньше общего случая. Этот факт известен как «правило трех сигма». Таким образом, для нормального распределения большие отклонения от мат.ожидания практически невозможны. Говорят, что нормальное распределение имеет очень легкие хвосты. Это обстоятельство необходимо учитывать при приближении случайных величин нормальными.

Случайная величина $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ имеет многомерное нормальное распределение с параметрами $\boldsymbol{\mu} \in \mathbb{R}^d$ и $\Sigma \in \mathbb{R}^{d\times d}$, если ее плотность задается выражением (см. рис. 1,б)

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi}^d\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

Параметр $\boldsymbol{\mu}$ является мат.ожиданием, а Σ — матрицей ковариации нормального распределения. Матрица Σ является симметричной и положительно определенной (в дальнейшем положительная определенность будет обозначаться как $\Sigma \succ 0$).

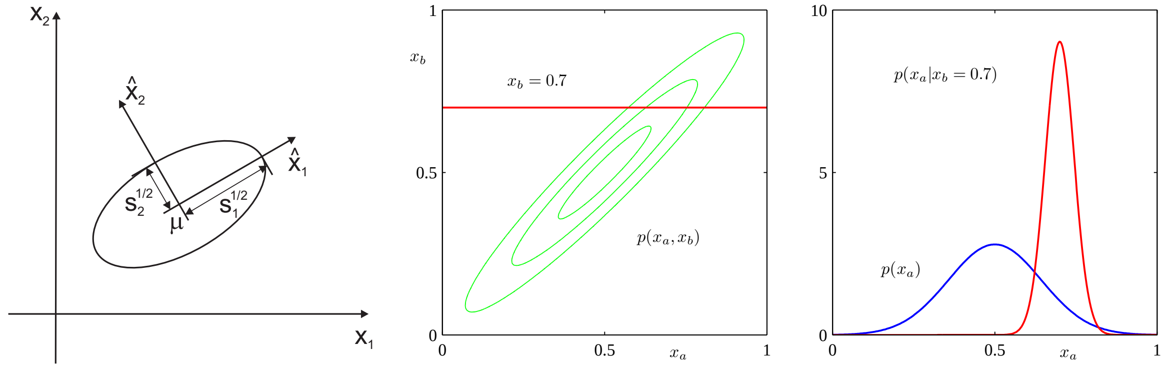
Нормальное распределение полностью определяется своими первым и вторым моментом, т.е. мат.ожиданием и матрицей ковариации. В частности, все зависимости между переменными x_i 3-го и более порядков (например, выражения вида $\mathbb{E}x_{i_1}x_{i_2}\ldots x_{i_m},\ m > 2$) являются функциями от зависимостей порядка ≤ 2.

Линии уровня плотности нормального распределения соответствуют линиям уровня квадратичной формы $(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ и представляют собой эллипсы. Можно выделить три основных вида этих линий уровня (см. рис. 2) в зависимости от значения матрицы ковариации. Если матрица ковариации пропорциональна единичной, т.е. имеет вид $\Sigma = \alpha I$, то все компоненты нормального распределения x_i являются независимыми друг от друга и имеют одинаковую дисперсию α. Линии уровня при этом образуют окружности. Диагональная матрица ковариации соответствует независимым компонентам x_i, но с различными дисперсиями. Линии уровня в этом случае являются эллипсами, параллельными координатным осям. Наконец, произвольная положительная определенная матрица ковариации соответствует эллипсам общего вида.

Матрица квадратичной формы $(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ является симметричной и положительно определенной. Поэтому с помощью ортогонального преобразования координат (поворота координатных осей) эту квадратичную форму можно привести к каноническому виду с положительными коэффициентами:

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \left\{\begin{array}{l}\hat{\boldsymbol{x}} = Q(\boldsymbol{x}-\boldsymbol{\mu})\\ Q^{-1} = Q^T\end{array}\right\} = \sum_i \frac{\hat{x}_i^2}{s_i}.$$

Здесь $s_i > 0$ — собственные значения матрицы Σ, а матрица поворота Q состоит из собственных векторов матрицы Σ. При этом значения $\sqrt{s_i}$ определяют длины полуосей эллипса, соответствующего линии уровня нормальной плотности $\exp(-1/2)/(\sqrt{2\pi}^d\sqrt{\det\Sigma})$, а базис в пространстве координат $\hat{\boldsymbol{x}}$ определяется собственными векторами матрицы Σ (см. рис. 3а).



(а) (б) (в)

Рис. 3: (а) — линия уровня нормальной плотности, соответствующая значению $\exp(-1/2)/(\sqrt{2\pi}^d\sqrt{\det\Sigma})$, (б) — линии уровня нормального распределения общего вида $p(x_a, x_b)$ в двухмерном пространстве, (в) — маргинальное распределение $p(x_a)$ (синяя кривая) и условное маргинальное распределение $p(x_a|x_b = 0.7)$ (красная кривая).

Разобьем вектор $\boldsymbol{x}$ на две группы переменных $\boldsymbol{x}_a, \boldsymbol{x}_b$ и обозначим

$$\boldsymbol{x} = \begin{bmatrix}\boldsymbol{x}_a\\ \boldsymbol{x}_b\end{bmatrix},\ \boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{bmatrix},\ \Sigma = \begin{bmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{bmatrix},\ \Lambda = \Sigma^{-1} = \begin{bmatrix}\Lambda_{aa} & \Lambda_{ab}\\ \Lambda_{ba} & \Lambda_{bb}\end{bmatrix}.$$

Матрицу Λ называют также матрицей точности. Тогда можно показать, что

$$\begin{aligned} p(\boldsymbol{x}_a) &= \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \Sigma_{aa}), \\ p(\boldsymbol{x}_a|\boldsymbol{x}_b) &= \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b), \Lambda_{aa}^{-1}). \end{aligned} \tag{1}$$

Этот результат означает, что вектор мат.ожиданий $\boldsymbol{\mu}$ состоит из мат.ожиданий отдельных компонент x_i, а на диагонали матрицы ковариации Σ стоят дисперсии соответствующих компонент x_i. Кроме того, у многомерного нормального распределения все маргинальные и условные распределения также являются нормальными (см. рис. 3б,в).

Рассмотрим величину $\boldsymbol{y} \in \mathbb{R}^D$, которая с точностью до нормального шума связана линейно с величиной $\boldsymbol{x}$, т.е.

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{y}|A\boldsymbol{x}, \Gamma), \quad A \in \mathbb{R}^{D\times d}, \Gamma \in \mathbb{R}^{D\times D}, \\ p(\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma), \quad \boldsymbol{\mu} \in \mathbb{R}^d,\ \Sigma \in \mathbb{R}^{d\times d}. \end{aligned}$$

Тогда можно показать, что

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|A\boldsymbol{\mu}, \Gamma + A\Sigma A^T), \tag{2}$$

$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}|P(A^T\Gamma^{-1}\boldsymbol{y} + \Sigma^{-1}\boldsymbol{\mu}), P),\ \ P = (\Sigma^{-1} + A^T\Gamma^{-1}A)^{-1}. \tag{3}$$

В частности, если $\Gamma = 0$, то результат (2) говорит о том, что любые линейные комбинации компонент нормального распределения также распределены нормально.

Найдем оценки максимального правдоподобия для параметров нормального распределения $\boldsymbol{\mu}, \Sigma$. Пусть имеется н.о.р. выборка объема N $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$, взятая из распределения $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$. Тогда оценки максимального правдоподобия являются решениями следующей задачи:

$$p(X|\boldsymbol{\mu}, \Sigma) = \prod_n \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}, \Sigma) \to \max_{\boldsymbol{\mu}, \Sigma}.$$

Переходя к логарифму, получаем:

$$\log p(X|\boldsymbol{\mu}, \Sigma) = -\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) - \frac{N}{2}\log\det\Sigma + \text{Const}.$$

Отсюда

$$\nabla_{\boldsymbol{\mu}} \log p(X|\boldsymbol{\mu},\Sigma) = -N\Sigma^{-1}\boldsymbol{\mu} - \sum_{n=1}^{N}\Sigma^{-1}\boldsymbol{x}_n = -\Sigma^{-1}\left(N\boldsymbol{\mu} - \sum_{n=1}^{N}\boldsymbol{x}_n\right) = 0 \quad \Rightarrow \quad \boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n.$$

Для получения оценки Σ_{ML} удобно перейти к матрице точности Λ:

$$\nabla_{\Lambda} \log p(X|\boldsymbol{\mu},\Sigma) = -\frac{1}{2}\nabla_{\Lambda}\mathrm{trace}\left(\Lambda\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T\right) + \frac{N}{2\det\Lambda}\nabla_{\Lambda}\det\Lambda =$$

$$-\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T + \frac{N}{2\det\Lambda}\det\Lambda\Lambda^{-1} = 0 \quad \Rightarrow \quad \Sigma_{ML} = \Lambda_{ML}^{-1} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T.$$