Численные методы проверки обоснованности обобщенных линейных моделей зависимостей

Левдик Павел Владимирович
Московский физико-технический институт

Моттль Вадим Вячеславович
Вычислительный центр РАН

Красоткина Ольга Вячеславовна
Московский государственный университет

Татарчук Александр Игоревич
Вычислительный центр РАН

Итак, мы рассматриваем вероятностную модель произвольной зависимости:

Априорная плотность распределения направляющего вектора и сдвига

$$\Psi(\boldsymbol{a},b|\mu) \propto \exp\left\{-V(\boldsymbol{a}|\mu)\right\}$$

Условная плотность распределения характеристик случайного объекта

$$\varphi(\boldsymbol{x},y\,|\,\boldsymbol{a},b,c) \propto \exp\left\{-cq\left(y,z(\boldsymbol{x},\boldsymbol{a},b)\right)\right\}$$

Итак, мы рассматриваем вероятностную модель произвольной зависимости:

Регуляризующая функция выбрана наблюдателем

$$\boxed{V(\boldsymbol{a}|\mu):\mathbb{R}^n \xrightarrow{\mu} \mathbb{R}^+} \qquad \Psi(\boldsymbol{a},b|\mu) \propto \exp\left\{-V(\boldsymbol{a}|\mu)\right\}$$

Функция связи выбрана наблюдателем

$$\boxed{q(y,z)} \qquad \varphi(\boldsymbol{x},y\,|\,\boldsymbol{a},b,c) \propto \exp\left\{-cq\left(y,z(\boldsymbol{x},\boldsymbol{a},b)\right)\right\}$$

Итак, мы рассматриваем вероятностную модель произвольной зависимости:

Регуляризующая функция выбрана наблюдателем

$$\boxed{V(\boldsymbol{a}|\mu):\mathbb{R}^n \xrightarrow{\mu} \mathbb{R}^+} \qquad \Psi(\boldsymbol{a},b|\mu) \propto \exp\{-V(\boldsymbol{a}|\mu)\}$$

Функция связи выбрана наблюдателем

$$\boxed{q(y,z)} \qquad \varphi(\boldsymbol{x},y\,|\,\boldsymbol{a},b,c) \propto \exp\{-cq\big(y,z(\boldsymbol{x},\boldsymbol{a},b)\big)\}$$

Все остальное стандартно:

Совместная условная плотность распределения случайной обучающей совокупности

$$\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c) = \prod_{j=1}^{N} \varphi(\boldsymbol{x}_j, y_j\,|\,\boldsymbol{a},b,c)$$

Структурные параметры обобщенной линейной модели зависимости

– параметр μ регуляр. функции,
– параметр c модели наблюдения.

Итак, мы рассматриваем вероятностную модель произвольной зависимости:

Регуляризующая функция выбрана наблюдателем

$$\boxed{V(\boldsymbol{a}|\mu):\mathbb{R}^n \xrightarrow{\mu} \mathbb{R}^+} \qquad \Psi(\boldsymbol{a},b|\mu) \propto \exp\{-V(\boldsymbol{a}|\mu)\}$$

Функция связи выбрана наблюдателем

$$\boxed{q(y,z)} \qquad \varphi(\boldsymbol{x},y|\boldsymbol{a},b,c) \propto \exp\{-cq(y,z(\boldsymbol{x},\boldsymbol{a},b))\}$$

Все остальное стандартно:

Совместная условная плотность распределения случайной обучающей совокупности

$$\Phi(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{a},b,c) = \prod_{j=1}^{N} \varphi(\boldsymbol{x}_j, y_j|\boldsymbol{a},b,c)$$

Структурные параметры обобщенной линейной модели зависимости

– параметр μ регуляр. функции,
– параметр c модели наблюдения.

Маргинальная плотность распределения обучающей совокупности – непрерывная смесь условных распределений:

$$F(\boldsymbol{X},\boldsymbol{Y}|\mu,c) = \int_{\mathbb{R}}\int_{\mathbb{R}^n} \Phi(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{a},b,c)\Psi(\boldsymbol{a},b|\mu)d\boldsymbol{a}db$$

Функция правдоподобия для (μ,c) (Marginal Likelihood, Evidence Function)

Оценки максимального правдоподобия для структурных параметров:

$$(\hat{\mu},\hat{c}) = \operatorname{argmax} F(\boldsymbol{X},\boldsymbol{Y}|\mu,c)$$

Итак, мы рассматриваем вероятностную модель произвольной зависимости:

Регуляризующая функция выбрана наблюдателем $\boxed{V(\boldsymbol{a}|\mu): \mathbb{R}^n \xrightarrow{\mu} \mathbb{R}^+}$ $\Psi(\boldsymbol{a},b|\mu) \propto \exp\{-V(\boldsymbol{a}|\mu)\}$

Функция связи выбрана наблюдателем $\boxed{q(y,z)}$ $\varphi(\boldsymbol{x},y|\boldsymbol{a},b,c) \propto \exp\{-cq(y,z(\boldsymbol{x},\boldsymbol{a},b))\}$

Все остальное стандартно:

Совместная условная плотность распределения случайной обучающей совокупности $\Phi(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{a},b,c) = \prod_{j=1}^{N} \varphi(\boldsymbol{x}_j, y_j|\boldsymbol{a},b,c)$

Структурные параметры обобщенной линейной модели зависимости – параметр μ регуляр. функции,
– параметр c модели наблюдения.

Маргинальная плотность распределения обучающей совокупности – непрерывная смесь условных распределений:

$F(\boldsymbol{X},\boldsymbol{Y}|\mu,c) = \int_{\mathbb{R}}\int_{\mathbb{R}^n} \Phi(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{a},b,c)\Psi(\boldsymbol{a},b|\mu)d\boldsymbol{a}db$ Функция правдоподобия для (μ,c) (Marginal Likelihood, Evidence Function)

Оценки максимального правдоподобия для структурных параметров:

$(\hat{\mu},\hat{c}) = \operatorname{argmax} F(\boldsymbol{X},\boldsymbol{Y}|\mu,c)$

Однако поиск точки максимума этой маргинальной плотности представляет собой крайне трудную вычислительную задачу.

Альтернативная запись функции правдоподобия

Два эквивалентных выражения для совместная плотность распределения наблюдений, направляющего вектора и сдвига:

$$\underbrace{H(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{a},b\,|\,\mu,c)}_{\substack{\textit{совместное}\\ \textit{распределение}}} = \underbrace{\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)}_{\substack{\textit{условное}\\ \textit{распределение}}} \underbrace{\Psi(\boldsymbol{a},b|\mu)}_{\substack{\textit{априорное}\\ \textit{распределение}\\ \textit{функция}\\ \textit{правдоподобия}}} = \underbrace{F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c)}_{\substack{\textit{маргинальное}\\ \textit{распределение}}} \underbrace{P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)}_{\substack{\textit{апостериорное}\\ \textit{распределение}}}$$

Альтернативная запись функции правдоподобия

Два эквивалентных выражения для совместная плотность распределения наблюдений, направляющего вектора и сдвига:

$$\underbrace{H(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{a},b\,|\,\mu,c)}_{\substack{\textit{совместное}\\ \textit{распределение}}} = \underbrace{\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)}_{\substack{\textit{условное}\\ \textit{распределение}}} \underbrace{\Psi(\boldsymbol{a},b|\mu)}_{\substack{\textit{априорное}\\ \textit{распределение}\\ \textit{функция}\\ \textit{правдоподобия}}} = \underbrace{F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c)}_{\substack{\textit{маргинальное}\\ \textit{распределение}}} \underbrace{P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)}_{\substack{\textit{апостериорное}\\ \textit{распределение}}}$$

Отсюда прямое выражение для функции правдоподобия:

При любых значениях $(\boldsymbol{a},b)$

$$F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c) = \frac{\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\Psi(\boldsymbol{a},b|\mu)}{P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)}$$

Альтернативная запись функции правдоподобия

Два эквивалентных выражения для совместная плотность распределения наблюдений, направляющего вектора и сдвига:

$$\underbrace{H(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{a},b\,|\,\mu,c)}_{\substack{\textit{совместное}\\ \textit{распределение}}} = \underbrace{\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)}_{\substack{\textit{условное}\\ \textit{распределение}}} \quad \underbrace{\Psi(\boldsymbol{a},b|\mu)}_{\substack{\textit{априорное}\\ \textit{распределение}\\ \textit{функция}\\ \textit{правдоподобия}}} = \underbrace{F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c)}_{\substack{\textit{маргинальное}\\ \textit{распределение}}} \underbrace{P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)}_{\substack{\textit{апостериорное}\\ \textit{распределение}}}$$

Отсюда прямое выражение для функции правдоподобия:

При любых значениях $(\boldsymbol{a},b)$

$$F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c) = \frac{\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\Psi(\boldsymbol{a},b|\mu)}{P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)}$$

Логарифмическая запись: Наиболее правдоподобные значения структурных параметров модели:

$$\ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c) = \ln\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c) + \ln\Psi(\boldsymbol{a},b|\mu) - \ln P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c) \to \max(\mu,c)$$

Альтернативная запись функции правдоподобия

Два эквивалентных выражения для совместная плотность распределения наблюдений, направляющего вектора и сдвига:

$$\underbrace{H(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{a},b\,|\,\mu,c)}_{\substack{\textit{совместное}\\ \textit{распределение}}} = \underbrace{\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)}_{\substack{\textit{условное}\\ \textit{распределение}}} \quad \underbrace{\Psi(\boldsymbol{a},b|\mu)}_{\substack{\textit{априорное}\\ \textit{распределение}\\ \textit{функция}\\ \textit{правдоподобия}}} = \underbrace{F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c)}_{\substack{\textit{маргинальное}\\ \textit{распределение}}} \underbrace{P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)}_{\substack{\textit{апостериорное}\\ \textit{распределение}}}$$

Отсюда прямое выражение для функции правдоподобия:

При любых значениях $(\boldsymbol{a},b)$

$$F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c) = \frac{\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\Psi(\boldsymbol{a},b|\mu)}{P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)}$$

Логарифмическая запись: Наиболее правдоподобные значения структурных параметров модели:

$$\ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c) = \ln\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c) + \ln\Psi(\boldsymbol{a},b|\mu) - \ln P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c) \to \max(\mu,c)$$

Это представление логарифмической функции правдоподобия лежит в основе известного EM-алгоритма ее максимизации (Expectation- maximization).

Нам нужно реализовать EM-алгоритм максимизации функции правдоподобия

Если бы мы могли вычислить апостериорную плотность для некоторых значений $(\hat{\mu}_k, \hat{c}_k)$, рассматриваемых как очередное приближение к точке максимума

$$\ln F(\boldsymbol{X},\boldsymbol{Y}\mid\mu,c)=\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n}\left[\ln\Phi(\boldsymbol{X},\boldsymbol{Y}\mid\boldsymbol{a},b,c)\right]P(\boldsymbol{a},b\mid\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)d\boldsymbol{a}db+$$

$$\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n}\left[\ln\Psi(\boldsymbol{a},b|\mu)\right]P(\boldsymbol{a},b\mid\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)d\boldsymbol{a}db-\underbrace{\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n}\ln P(\boldsymbol{a},b\mid\boldsymbol{X},\boldsymbol{Y},\mu,c)P(\boldsymbol{a},b\mid\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)d\boldsymbol{a}db},$$

то выполнялось бы фундаментальное неравенство:

$$\leq\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n}\ln P(\boldsymbol{a},b\mid\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)P(\boldsymbol{a},b\mid\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)d\boldsymbol{a}db$$

Нам нужно реализовать EM-алгоритм максимизации функции правдоподобия

Если бы мы могли вычислить апостериорную плотность для некоторых значений $(\hat{\mu}_k, \hat{c}_k)$, рассматриваемых как очередное приближение к точке максимума

$$\ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c) = \int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n} \left[\ln \Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\right] P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}db +$$

$$\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n} \left[\ln \Psi(\boldsymbol{a},b|\mu)\right] P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}db - \underbrace{\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n} \ln P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c) P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}db}_{},$$

то выполнялось бы фундаментальное неравенство:

$$\leq \int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n} \ln P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k) P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}db$$

Отсюда следует правило пересчета $(\hat{\mu}_k, \hat{c}_k) \to (\hat{\mu}_{k+1}, \hat{c}_{k+1})$, гарантирующее неубывание функции правдоподобия (EM- алгоритм):

$$\hat{\mu}_{k+1} = \arg\max_{\mu} \int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n} \left[\ln \Psi(\boldsymbol{a},b|\mu)\right] P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}db,$$

$$\hat{c}_{k+1} = \underbrace{\arg\max_{c}}_{\text{maximization}} \underbrace{\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n} \left[\ln \Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\right] P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}db}_{\text{expectation}}.$$

Нам нужно реализовать EM-алгоритм максимизации функции правдоподобия

Если бы мы могли вычислить апостериорную плотность для некоторых значений $(\hat{\mu}_k, \hat{c}_k)$, рассматриваемых как очередное приближение к точке максимума

$$\ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c) = \int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n} \left[\ln \Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\right] P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}db +$$

$$\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n} \left[\ln \Psi(\boldsymbol{a},b|\mu)\right] P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}db - \underbrace{\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n} \ln P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c) P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}db}_{},$$

то выполнялось бы фундаментальное неравенство:

$$\leq \int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n} \ln P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k) P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}db$$

Отсюда следует правило пересчета $(\hat{\mu}_k, \hat{c}_k) \to (\hat{\mu}_{k+1}, \hat{c}_{k+1})$, гарантирующее неубывание функции правдоподобия (EM- алгоритм):

$$\hat{\mu}_{k+1} = \arg\max_{\mu} \int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n} \left[\ln \Psi(\boldsymbol{a},b|\mu)\right] P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}db,$$

$$\hat{c}_{k+1} = \underbrace{\arg\max_{c}}_{\text{maximization}} \underbrace{\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}^n} \left[\ln \Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\right] P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\,d\boldsymbol{a}db}_{\text{expectation}}.$$

Теорема: $\ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\hat{\mu}_{k+1},\hat{c}_{k+1}) \geq \ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\hat{\mu}_k,\hat{c}_k)$,

пока $\dfrac{\partial}{\partial\mu}\ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\hat{\mu}_k,\hat{c}_k) \neq 0$ либо $\dfrac{\partial}{\partial c}\ln F(\boldsymbol{X},\boldsymbol{Y}\,|\,\hat{\mu}_k,\hat{c}_k) \neq 0$.

Численно реализовать исходный EM-алгоритм невозможно

Еще раз EM-алгоритм:

$$\hat{\mu}_{k+1} = \arg\max_{\mu} \int_{\mathbb{R}} \int_{\mathbb{R}^n} \left[\ln \Psi(\boldsymbol{a}, b|\mu)\right] P(\boldsymbol{a}, b \mid \boldsymbol{X}, \boldsymbol{Y}, \hat{\mu}_k, \hat{c}_k) d\boldsymbol{a} db,$$

$$\hat{c}_{k+1} = \arg\max_{\mu} \int_{\mathbb{R}} \int_{\mathbb{R}^n} \left[\ln \Phi(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{a}, b, c)\right] P(\boldsymbol{a}, b \mid \boldsymbol{X}, \boldsymbol{Y}, \hat{\mu}_k, \hat{c}_k) d\boldsymbol{a} db.$$

Численно реализовать исходный EM-алгоритм невозможно

Еще раз EM-алгоритм:

$$\hat{\mu}_{k+1} = \arg\max_{\mu} \int_{\mathbb{R}} \int_{\mathbb{R}^n} \left[\ln \Psi(\boldsymbol{a}, b|\mu)\right] P(\boldsymbol{a}, b \mid \boldsymbol{X}, \boldsymbol{Y}, \hat{\mu}_k, \hat{c}_k) d\boldsymbol{a} db,$$

$$\hat{c}_{k+1} = \arg\max_{\mu} \int_{\mathbb{R}} \int_{\mathbb{R}^n} \left[\ln \Phi(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{a}, b, c)\right] P(\boldsymbol{a}, b \mid \boldsymbol{X}, \boldsymbol{Y}, \hat{\mu}_k, \hat{c}_k) d\boldsymbol{a} db.$$

Апостериорное распределение направляющего вектора и сдвига:

$$P(\boldsymbol{a}, b \mid \boldsymbol{X}, \boldsymbol{Y}, \mu, c) \propto \Psi(\boldsymbol{a}, b|\mu)\Phi(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{a}, b, c)$$

Численно реализовать исходный EM-алгоритм невозможно

Еще раз EM-алгоритм:

$$\hat{\mu}_{k+1} = \arg\max_{\mu} \int_{\mathbb{R}} \int_{\mathbb{R}^n} \left[\ln \Psi(\boldsymbol{a}, b|\mu)\right] P(\boldsymbol{a}, b \mid \boldsymbol{X}, \boldsymbol{Y}, \hat{\mu}_k, \hat{c}_k) d\boldsymbol{a} db,$$

$$\hat{c}_{k+1} = \arg\max_{\mu} \int_{\mathbb{R}} \int_{\mathbb{R}^n} \left[\ln \Phi(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{a}, b, c)\right] P(\boldsymbol{a}, b \mid \boldsymbol{X}, \boldsymbol{Y}, \hat{\mu}_k, \hat{c}_k) d\boldsymbol{a} db.$$

Апостериорное распределение направляющего вектора и сдвига:

$$P(\boldsymbol{a}, b \mid \boldsymbol{X}, \boldsymbol{Y}, \mu, c) \propto \Psi(\boldsymbol{a}, b|\mu) \Phi(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{a}, b, c)$$

$$\Psi(\boldsymbol{a}, b|\mu) \propto \exp\left\{-V(\boldsymbol{a}|\mu)\right\}, \quad \Phi(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{a}, b, c) \propto \exp\left\{-c \sum_{j=1}^{N} q\left(y_j, z(\boldsymbol{x}_j, \boldsymbol{a}, b)\right)\right\}$$

Для многих регуляризующих функций $V(\boldsymbol{a}|\mu)$ и функций связи $q(y, z)$ интегрирование по такой апостериорной плотности затруднительно.

Численно реализовать исходный EM-алгоритм невозможно

Еще раз EM-алгоритм:

$$\hat{\mu}_{k+1} = \arg\max_{\mu} \int_{\mathbb{R}} \int_{\mathbb{R}^n} \left[\ln \Psi(\boldsymbol{a}, b|\mu)\right] P(\boldsymbol{a}, b \mid \boldsymbol{X}, \boldsymbol{Y}, \hat{\mu}_k, \hat{c}_k) d\boldsymbol{a} db,$$

$$\hat{c}_{k+1} = \arg\max_{\mu} \int_{\mathbb{R}} \int_{\mathbb{R}^n} \left[\ln \Phi(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{a}, b, c)\right] P(\boldsymbol{a}, b \mid \boldsymbol{X}, \boldsymbol{Y}, \hat{\mu}_k, \hat{c}_k) d\boldsymbol{a} db.$$

Апостериорное распределение направляющего вектора и сдвига:

$$P(\boldsymbol{a}, b \mid \boldsymbol{X}, \boldsymbol{Y}, \mu, c) \propto \Psi(\boldsymbol{a}, b|\mu) \Phi(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{a}, b, c)$$

$$\Psi(\boldsymbol{a}, b|\mu) \propto \exp\left\{-V(\boldsymbol{a}|\mu)\right\}, \quad \Phi(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{a}, b, c) \propto \exp\left\{-c \sum_{j=1}^{N} q\left(y_j, z(\boldsymbol{x}_j, \boldsymbol{a}, b)\right)\right\}$$

Для многих регуляризующих функций $V(\boldsymbol{a}|\mu)$ и функций связи $q(y, z)$ интегрирование по такой апостериорной плотности затруднительно.

Идея: Нормальная аппроксимация апостериорной плотности

$$P(\boldsymbol{a}, b \mid \boldsymbol{X}, \boldsymbol{Y}, \mu, c) \cong \underbrace{\mathcal{N}(\boldsymbol{a} \mid \hat{\boldsymbol{a}}_{\boldsymbol{y}, \boldsymbol{X}, \mu, c}, \boldsymbol{B}_{\boldsymbol{X}, c})}_{\substack{\textit{маргинальное} \\ \textit{апостеориорное} \\ \textit{распределение}}} \underbrace{\mathcal{N}(b \mid \hat{b}_{\boldsymbol{y}, \boldsymbol{X}, \mu, c}(\boldsymbol{a}), \hat{\sigma}_N^2)}_{\substack{\textit{условное} \\ \textit{апостеориорное} \\ \textit{распределение}}}.$$

Численно реализовать исходный EM-алгоритм невозможно

Еще раз EM-алгоритм:

$$\hat{\mu}_{k+1} = \arg\max_{\mu} \int_{\mathbb{R}} \int_{\mathbb{R}^n} \left[\ln \Psi(\boldsymbol{a},b|\mu)\right] P(\boldsymbol{a},b \mid \boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\, d\boldsymbol{a}\, db,$$

$$\hat{c}_{k+1} = \arg\max_{\mu} \int_{\mathbb{R}} \int_{\mathbb{R}^n} \left[\ln \Phi(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{a},b,c)\right] P(\boldsymbol{a},b \mid \boldsymbol{X},\boldsymbol{Y},\hat{\mu}_k,\hat{c}_k)\, d\boldsymbol{a}\, db.$$

Апостериорное распределение направляющего вектора и сдвига:

$$P(\boldsymbol{a},b \mid \boldsymbol{X},\boldsymbol{Y},\mu,c) \propto \Psi(\boldsymbol{a},b|\mu)\Phi(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{a},b,c)$$

$$\Psi(\boldsymbol{a},b|\mu) \propto \exp\left\{-V(\boldsymbol{a}|\mu)\right\}, \quad \Phi(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{a},b,c) \propto \exp\left\{-c\sum_{j=1}^{N} q\left(y_j, z(\boldsymbol{x}_j,\boldsymbol{a},b)\right)\right\}$$

Для многих регуляризующих функций $V(\boldsymbol{a}|\mu)$ и функций связи $q(y,z)$ интегрирование по такой апостериорной плотности затруднительно.

Идея: Нормальная аппроксимация апостериорной плотности

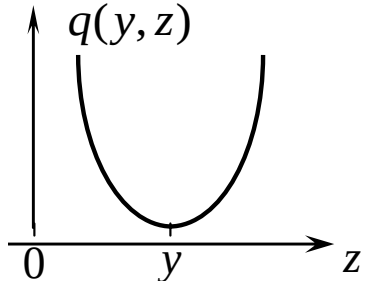
$$P(\boldsymbol{a},b \mid \boldsymbol{X},\boldsymbol{Y},\mu,c) \cong \underbrace{\mathcal{N}(\boldsymbol{a} \mid \hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c}, \boldsymbol{B}_{\boldsymbol{X},c})}_{\substack{\textit{маргинальное} \\ \textit{апостеориорное} \\ \textit{распределение}}} \underbrace{\mathcal{N}(b \mid \hat{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\boldsymbol{a}), \hat{\sigma}_N^2)}_{\substack{\textit{условное} \\ \textit{апостеориорное} \\ \textit{распределение}}}.$$

Ограничимся задачей восстановления числовой регрессии с квадратично-модульной регуляризацией

Регуляризация Elastic Net $\quad \boldsymbol{a} \in \mathbb{R}^n$

$$V(\boldsymbol{a} \mid \mu) = \sum_{i=1}^{n} a_i^2 + \mu \sum_{i=1}^{n} |a_i|$$

Квадратичная функция связи



Оценивание параметров регрессии: Минимум регуляризованного эмпирического риска

Обучающая совокупность $\left\{(\boldsymbol{x}_j, y_j),\ j=1,...,N\right\},\quad \boldsymbol{x}_j \in \mathbb{R}^n,\ y_j \in \mathbb{R}$

$$(\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c}, \hat{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}) = \arg\min_{\boldsymbol{a},b} \left\{ \sum_{i=1}^{n} \left(a_i^2 + \mu |a_i| \right) + \frac{1}{2c} \sum_{j=1}^{N} \left[y_j - \left(\sum_{i=1}^{n} x_{ij} a_i + b \right) \right]^2 \right\}$$

Оценивание параметров регрессии: Минимум регуляризованного эмпирического риска

Обучающая совокупность $\left\{(\boldsymbol{x}_j, y_j),\ j=1,...,N\right\},\quad \boldsymbol{x}_j \in \mathbb{R}^n,\ y_j \in \mathbb{R}$

$$(\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c}, \hat{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}) = \arg\min_{\boldsymbol{a},b} \left\{ \sum_{i=1}^{n} \left(a_i^2 + \mu |a_i| \right) + \frac{1}{2c} \sum_{j=1}^{N} \left[y_j - \left(\sum_{i=1}^{n} x_{ij} a_i + b \right) \right]^2 \right\}$$

Теорема: Пусть $(\hat{\delta}_1,...,\hat{\delta}_N)$ есть решение задачи выпуклой оптимизации

$$W_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\delta_1,...,\delta_N) = \sum_{i=1}^{n} \left\{ \min\left[\frac{\mu}{2} + \sum_{j=1}^{N} \delta_j x_{ij},\ 0,\ \frac{\mu}{2} - \sum_{j=1}^{N} \delta_j x_{ij} \right] \right\}^2 + 2c \sum_{j=1}^{N} (\delta_j - y_j)^2 \to \begin{cases} \min(\delta_1,...,\delta_N), \\ \sum_{j=1}^{N} \delta_j = 0. \end{cases}$$

Оценивание параметров регрессии: Минимум регуляризованного эмпирического риска

Обучающая совокупность $\left\{(\boldsymbol{x}_j, y_j),\ j=1,...,N\right\},\quad \boldsymbol{x}_j \in \mathbb{R}^n,\ y_j \in \mathbb{R}$

$$(\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c}, \hat{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}) = \arg\min_{\boldsymbol{a},b} \left\{ \sum_{i=1}^{n} \left(a_i^2 + \mu |a_i| \right) + \frac{1}{2c} \sum_{j=1}^{N} \left[y_j - \left(\sum_{i=1}^{n} x_{ij} a_i + b \right) \right]^2 \right\}$$

Теорема: Пусть $(\hat{\delta}_1,...,\hat{\delta}_N)$ есть решение задачи выпуклой оптимизации

$$W_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\delta_1,...,\delta_N) = \sum_{i=1}^{n} \left\{ \min\left[\frac{\mu}{2} + \sum_{j=1}^{N} \delta_j x_{ij},\ 0,\ \frac{\mu}{2} - \sum_{j=1}^{N} \delta_j x_{ij} \right] \right\}^2 + 2c \sum_{j=1}^{N} (\delta_j - y_j)^2 \to \begin{cases} \min(\delta_1,...,\delta_N), \\ \sum_{j=1}^{N} \delta_j = 0. \end{cases}$$

Тогда оценки параметров регрессии определяются выражениями:

$$\hat{a}_{i,\boldsymbol{y},\boldsymbol{X},\mu,c} = \frac{1}{\beta} \begin{cases} \left(\sum_{j=1}^{N} \hat{\delta}_j x_{ij} + \mu \right) < 0, & \sum_{j=1}^{N} \hat{\delta}_j x_{ij} < -\frac{\mu}{2}, \\ 0, & -\frac{\mu}{2} \le \sum_{j=1}^{N} \hat{\delta}_j x_{ij} \le \frac{\mu}{2}, \\ \left(\sum_{j=1}^{N} \hat{\delta}_j x_{ij} - \mu \right) > 0, & \sum_{j=1}^{N} \hat{\delta}_j x_{ij} > \frac{\mu}{2}, \end{cases} \qquad \begin{array}{l} i = 1,...,n, \\ \hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c} = (\hat{a}_{1,\boldsymbol{y},\boldsymbol{X},\mu,c} \cdots \hat{a}_{n,\boldsymbol{y},\boldsymbol{X},\mu,c})^T \in \mathbb{R}^n, \end{array}$$

$$\hat{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c} = \frac{1}{N} \sum_{j=1}^{N} \left(y_j - \sum_{i=1}^{n} x_{ij} \hat{a}_{i,\boldsymbol{y},\boldsymbol{X},\mu,c} \right).$$

Оценивание параметров регрессии: Минимум регуляризованного эмпирического риска

Обучающая совокупность $\left\{(\boldsymbol{x}_j, y_j),\ j=1,...,N\right\},\quad \boldsymbol{x}_j \in \mathbb{R}^n,\ y_j \in \mathbb{R}$

$$(\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c}, \hat{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}) = \arg\min_{\boldsymbol{a},b} \left\{ \sum_{i=1}^{n} \left(a_i^2 + \mu |a_i| \right) + \frac{1}{2c} \sum_{j=1}^{N} \left[y_j - \left(\sum_{i=1}^{n} x_{ij} a_i + b \right) \right]^2 \right\}$$

Теорема: Пусть $(\hat{\delta}_1,...,\hat{\delta}_N)$ есть решение задачи выпуклой оптимизации

$$W_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\delta_1,...,\delta_N) = \sum_{i=1}^{n} \left\{ \min\left[\frac{\mu}{2} + \sum_{j=1}^{N} \delta_j x_{ij},\ 0,\ \frac{\mu}{2} - \sum_{j=1}^{N} \delta_j x_{ij} \right] \right\}^2 + 2c \sum_{j=1}^{N} (\delta_j - y_j)^2 \to \begin{cases} \min(\delta_1,...,\delta_N), \\ \sum_{j=1}^{N} \delta_j = 0. \end{cases}$$

Тогда оценки параметров регрессии определяются выражениями:

$$\hat{a}_{i,\boldsymbol{y},\boldsymbol{X},\mu,c} = \frac{1}{\beta} \begin{cases} \left(\sum_{j=1}^{N} \hat{\delta}_j x_{ij} + \mu \right) < 0, & \sum_{j=1}^{N} \hat{\delta}_j x_{ij} < -\frac{\mu}{2}, \\ 0, & -\frac{\mu}{2} \le \sum_{j=1}^{N} \hat{\delta}_j x_{ij} \le \frac{\mu}{2}, \\ \left(\sum_{j=1}^{N} \hat{\delta}_j x_{ij} - \mu \right) > 0, & \sum_{j=1}^{N} \hat{\delta}_j x_{ij} > \frac{\mu}{2}, \end{cases} \qquad \begin{array}{l} i = 1,...,n, \\ \hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c} = (\hat{a}_{1,\boldsymbol{y},\boldsymbol{X},\mu,c} \cdots \hat{a}_{n,\boldsymbol{y},\boldsymbol{X},\mu,c})^T \in \mathbb{R}^n, \end{array}$$

$$\hat{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c} = \frac{1}{N} \sum_{j=1}^{N} \left(y_j - \sum_{i=1}^{n} x_{ij} \hat{a}_{i,\boldsymbol{y},\boldsymbol{X},\mu,c} \right).$$

Очевидный эффект автоматического отбора признаков.

Оценивание параметров регрессии: Минимум регуляризованного эмпирического риска

Обучающая совокупность $\left\{(\boldsymbol{x}_j, y_j),\ j=1,...,N\right\},\quad \boldsymbol{x}_j \in \mathbb{R}^n,\ y_j \in \mathbb{R}$

$$(\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c}, \hat{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}) = \arg\min_{\boldsymbol{a},b} \left\{ \sum_{i=1}^{n} \left(a_i^2 + \mu |a_i| \right) + \frac{1}{2c} \sum_{j=1}^{N} \left[y_j - \left(\sum_{i=1}^{n} x_{ij} a_i + b \right) \right]^2 \right\}$$

Теорема: Пусть $(\hat{\delta}_1,...,\hat{\delta}_N)$ есть решение задачи выпуклой оптимизации

$$W_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\delta_1,...,\delta_N) = \sum_{i=1}^{n} \left\{ \min\left[\frac{\mu}{2} + \sum_{j=1}^{N} \delta_j x_{ij},\ 0,\ \frac{\mu}{2} - \sum_{j=1}^{N} \delta_j x_{ij} \right] \right\}^2 + 2c \sum_{j=1}^{N} (\delta_j - y_j)^2 \to \begin{cases} \min(\delta_1,...,\delta_N), \\ \sum_{j=1}^{N} \delta_j = 0. \end{cases}$$

Тогда оценки параметров регрессии определяются выражениями:

$$\hat{a}_{i,\boldsymbol{y},\boldsymbol{X},\mu,c} = \frac{1}{\beta} \begin{cases} \left(\sum_{j=1}^{N} \hat{\delta}_j x_{ij} + \mu \right) < 0, & \sum_{j=1}^{N} \hat{\delta}_j x_{ij} < -\frac{\mu}{2}, \\ 0, & -\frac{\mu}{2} \le \sum_{j=1}^{N} \hat{\delta}_j x_{ij} \le \frac{\mu}{2}, \\ \left(\sum_{j=1}^{N} \hat{\delta}_j x_{ij} - \mu \right) > 0, & \sum_{j=1}^{N} \hat{\delta}_j x_{ij} > \frac{\mu}{2}, \end{cases} \qquad \begin{aligned} & i = 1,...,n, \\ & \hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c} = (\hat{a}_{1,\boldsymbol{y},\boldsymbol{X},\mu,c} \cdots \hat{a}_{n,\boldsymbol{y},\boldsymbol{X},\mu,c})^T \in \mathbb{R}^n, \end{aligned}$$

$$\hat{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c} = \frac{1}{N} \sum_{j=1}^{N} \left(y_j - \sum_{i=1}^{n} x_{ij} \hat{a}_{i,\boldsymbol{y},\boldsymbol{X},\mu,c} \right).$$

Очевидный эффект автоматического отбора признаков.

Три подмножества компонент вектора коэффициентов регрессии:

$$\hat{\mathbb{I}}^{+}_{\boldsymbol{y},\boldsymbol{X},\mu,c} = \left\{ i : \hat{a}_{i,\boldsymbol{y},\boldsymbol{X},\mu,c} > 0 \right\}, \hat{\mathbb{I}}^{-}_{\boldsymbol{y},\boldsymbol{X},\mu,c} = \left\{ i : \hat{a}_{i,\boldsymbol{y},\boldsymbol{X},\mu,c} < 0 \right\}$$

$$\hat{\mathbb{I}}^{0}_{\boldsymbol{y},\boldsymbol{X},\mu,c} = \left\{ i : \hat{a}_{i,\boldsymbol{y},\boldsymbol{X},\mu,c} = 0 \right\}$$

Оценивание параметров регрессии: Минимум регуляризованного эмпирического риска

Обучающая совокупность $\left\{(\boldsymbol{x}_j, y_j),\ j=1,...,N\right\}, \quad \boldsymbol{x}_j \in \mathbb{R}^n,\ y_j \in \mathbb{R}$

$$(\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c}, \hat{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}) = \arg\min_{\boldsymbol{a},b} \left\{ \sum_{i=1}^{n} \left(a_i^2 + \mu |a_i| \right) + \frac{1}{2c} \sum_{j=1}^{N} \left[y_j - \left(\sum_{i=1}^{n} x_{ij} a_i + b \right) \right]^2 \right\}$$

Теорема: Пусть $(\hat{\delta}_1,...,\hat{\delta}_N)$ есть решение задачи выпуклой оптимизации

$$W_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\delta_1,...,\delta_N) = \sum_{i=1}^{n} \left\{ \min\left[\frac{\mu}{2} + \sum_{j=1}^{N} \delta_j x_{ij},\ 0,\ \frac{\mu}{2} - \sum_{j=1}^{N} \delta_j x_{ij} \right] \right\}^2 + 2c \sum_{j=1}^{N} (\delta_j - y_j)^2 \to \begin{cases} \min(\delta_1,...,\delta_N), \\ \sum_{j=1}^{N} \delta_j = 0. \end{cases}$$

Тогда оценки параметров регрессии определяются выражениями:

$$\hat{a}_{i,\boldsymbol{y},\boldsymbol{X},\mu,c} = \frac{1}{\beta} \begin{cases} \left(\sum_{j=1}^{N} \hat{\delta}_j x_{ij} + \mu \right) < 0, & \sum_{j=1}^{N} \hat{\delta}_j x_{ij} < -\frac{\mu}{2}, \\ 0, & -\frac{\mu}{2} \le \sum_{j=1}^{N} \hat{\delta}_j x_{ij} \le \frac{\mu}{2}, \\ \left(\sum_{j=1}^{N} \hat{\delta}_j x_{ij} - \mu \right) > 0, & \sum_{j=1}^{N} \hat{\delta}_j x_{ij} > \frac{\mu}{2}, \end{cases} \qquad \begin{array}{l} i = 1,...,n, \\ \hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c} = (\hat{a}_{1,\boldsymbol{y},\boldsymbol{X},\mu,c} \cdots \hat{a}_{n,\boldsymbol{y},\boldsymbol{X},\mu,c})^T \in \mathbb{R}^n, \end{array}$$

$$\hat{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c} = \frac{1}{N} \sum_{j=1}^{N} \left(y_j - \sum_{i=1}^{n} x_{ij} \hat{a}_{i,\boldsymbol{y},\boldsymbol{X},\mu,c} \right).$$

Очевидный эффект автоматического отбора признаков.

Чем больше параметр селективности $\mu \ge 0$, тем меньше ненулевых коэффициентов регрессии, т.е. меньше уровень сложности модели.

Нормальная аппроксимация апостериорной плотности

Исходная «ненормальная» апостериорная плотность случайных параметров регрессии:

$$P(\boldsymbol{a},b\,|\,\boldsymbol{y},\boldsymbol{X},\mu,c)\propto\Psi(\boldsymbol{a}\,|\,\mu)\Phi(\boldsymbol{y}\,|\,\boldsymbol{X},\boldsymbol{a},b,c)\propto$$

$$\exp\left\{-\left\{\boldsymbol{a}^T\boldsymbol{a}+\mu\sum_{i\in\hat{\mathbb{I}}^0_{\boldsymbol{y},\boldsymbol{X},\mu,c}}|a_i|+\mu\sum_{i\in\hat{\mathbb{I}}^+_{\boldsymbol{y},\boldsymbol{X},\mu,c}}|a_i|+\mu\sum_{i\in\hat{\mathbb{I}}^-_{\boldsymbol{y},\boldsymbol{X},\mu,c}}|a_i|+\frac{1}{2c}\left[\boldsymbol{y}-\left(\boldsymbol{X}^T\boldsymbol{a}+\boldsymbol{1}b\right)\right]^T\left[\boldsymbol{y}-\left(\boldsymbol{X}^T\boldsymbol{a}+\boldsymbol{1}b\right)\right]\right\}\right\}.$$

Нормальная аппроксимация апостериорной плотности

Исходная «ненормальная» апостериорная плотность случайных параметров регрессии:

$$P(\boldsymbol{a},b\,|\,\boldsymbol{y},\boldsymbol{X},\mu,c)\propto\Psi(\boldsymbol{a}\,|\,\mu)\Phi(\boldsymbol{y}\,|\,\boldsymbol{X},\boldsymbol{a},b,c)\propto$$

$$\exp\left\{-\left\{\boldsymbol{a}^T\boldsymbol{a}+\mu\underbrace{\sum_{i\in\hat{\mathbb{I}}^0_{\boldsymbol{y},\boldsymbol{X},\mu,c}}|a_i|}+\mu\underbrace{\sum_{i\in\hat{\mathbb{I}}^+_{\boldsymbol{y},\boldsymbol{X},\mu,c}}|a_i|}+\mu\underbrace{\sum_{i\in\hat{\mathbb{I}}^-_{\boldsymbol{y},\boldsymbol{X},\mu,c}}|a_i|}+\frac{1}{2c}\left[\boldsymbol{y}-\left(\boldsymbol{X}^T\boldsymbol{a}+\boldsymbol{1}b\right)\right]^T\left[\boldsymbol{y}-\left(\boldsymbol{X}^T\boldsymbol{a}+\boldsymbol{1}b\right)\right]\right\}\right\}.$$

Нормальная аппроксимация:

$$\hat{P}(\boldsymbol{a},b\,|\,\boldsymbol{y},\boldsymbol{X},\mu,c)=\mathcal{N}(\boldsymbol{a}\,|\,\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c},\boldsymbol{B}_{\boldsymbol{X},c})\mathcal{N}(b\,|\,\tilde{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\boldsymbol{a}),\tilde{\sigma}^2_N)\propto$$

$$\exp\left\{-\left\{\boldsymbol{a}^T\boldsymbol{a}+\mu\sum_{i\in\hat{\mathbb{I}}^+_{\boldsymbol{y},\boldsymbol{X},\mu,c}}a_i-\mu\sum_{i\in\hat{\mathbb{I}}^-_{\boldsymbol{y},\boldsymbol{X},\mu,c}}a_i+\frac{1}{2c}\left[\boldsymbol{y}-\left(\boldsymbol{X}^T\boldsymbol{a}+\boldsymbol{1}b\right)\right]^T\left[\boldsymbol{y}-\left(\boldsymbol{X}^T\boldsymbol{a}+\boldsymbol{1}b\right)\right]\right\}\right\}.$$

Нормальная аппроксимация апостериорной плотности

Исходная «ненормальная» апостериорная плотность случайных параметров регрессии:

$$P(\boldsymbol{a},b\,|\,\boldsymbol{y},\boldsymbol{X},\mu,c)\propto\Psi(\boldsymbol{a}\,|\,\mu)\Phi(\boldsymbol{y}\,|\,\boldsymbol{X},\boldsymbol{a},b,c)\propto$$

$$\exp\left\{-\left\{\boldsymbol{a}^T\boldsymbol{a}+\mu\sum_{i\in\hat{\mathbb{I}}^0_{\boldsymbol{y},\boldsymbol{X},\mu,c}}|a_i|+\mu\sum_{i\in\hat{\mathbb{I}}^+_{\boldsymbol{y},\boldsymbol{X},\mu,c}}|a_i|+\mu\sum_{i\in\hat{\mathbb{I}}^-_{\boldsymbol{y},\boldsymbol{X},\mu,c}}|a_i|+\frac{1}{2c}\left[\boldsymbol{y}-\left(\boldsymbol{X}^T\boldsymbol{a}+\boldsymbol{1}b\right)\right]^T\left[\boldsymbol{y}-\left(\boldsymbol{X}^T\boldsymbol{a}+\boldsymbol{1}b\right)\right]\right\}\right\}.$$

Нормальная аппроксимация:

$$\hat{P}(\boldsymbol{a},b\,|\,\boldsymbol{y},\boldsymbol{X},\mu,c)=\mathcal{N}(\boldsymbol{a}\,|\,\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c},\boldsymbol{B}_{\boldsymbol{X},c})\mathcal{N}(b\,|\,\tilde{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\boldsymbol{a}),\tilde{\sigma}^2_N)\propto$$

$$\exp\left\{-\left\{\boldsymbol{a}^T\boldsymbol{a}+\mu\sum_{i\in\hat{\mathbb{I}}^+_{\boldsymbol{y},\boldsymbol{X},\mu,c}}a_i-\mu\sum_{i\in\hat{\mathbb{I}}^-_{\boldsymbol{y},\boldsymbol{X},\mu,c}}a_i+\frac{1}{2c}\left[\boldsymbol{y}-\left(\boldsymbol{X}^T\boldsymbol{a}+\boldsymbol{1}b\right)\right]^T\left[\boldsymbol{y}-\left(\boldsymbol{X}^T\boldsymbol{a}+\boldsymbol{1}b\right)\right]\right\}\right\}.$$

Нормальная аппроксимация апостериорной плотности

Исходная «ненормальная» апостериорная плотность случайных параметров регрессии:

$$P(\boldsymbol{a},b\,|\,\boldsymbol{y},\boldsymbol{X},\mu,c)\propto \Psi(\boldsymbol{a}\,|\,\mu)\Phi(\boldsymbol{y}\,|\,\boldsymbol{X},\boldsymbol{a},b,c)\propto$$

$$\exp\left\{-\left\{\boldsymbol{a}^T\boldsymbol{a}+\mu\sum_{i\in\hat{\mathbb{I}}^0_{\boldsymbol{y},\boldsymbol{X},\mu,c}}|a_i|+\mu\sum_{i\in\hat{\mathbb{I}}^+_{\boldsymbol{y},\boldsymbol{X},\mu,c}}|a_i|+\mu\sum_{i\in\hat{\mathbb{I}}^-_{\boldsymbol{y},\boldsymbol{X},\mu,c}}|a_i|+\frac{1}{2c}\left[\boldsymbol{y}-\left(\boldsymbol{X}^T\boldsymbol{a}+\boldsymbol{1}b\right)\right]^T\left[\boldsymbol{y}-\left(\boldsymbol{X}^T\boldsymbol{a}+\boldsymbol{1}b\right)\right]\right\}\right\}.$$

Нормальная аппроксимация:

$$\hat{P}(\boldsymbol{a},b\,|\,\boldsymbol{y},\boldsymbol{X},\mu,c)=\mathcal{N}(\boldsymbol{a}\,|\,\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c},\boldsymbol{B}_{\boldsymbol{X},c})\mathcal{N}(b\,|\,\tilde{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\boldsymbol{a}),\tilde{\sigma}^2_N)\propto$$

$$\exp\left\{-\left\{\boldsymbol{a}^T\boldsymbol{a}\quad+\mu\sum_{i\in\hat{\mathbb{I}}^+_{\boldsymbol{y},\boldsymbol{X},\mu,c}}a_i-\mu\sum_{i\in\hat{\mathbb{I}}^-_{\boldsymbol{y},\boldsymbol{X},\mu,c}}a_i+\frac{1}{2c}\left[\boldsymbol{y}-\left(\boldsymbol{X}^T\boldsymbol{a}+\boldsymbol{1}b\right)\right]^T\left[\boldsymbol{y}-\left(\boldsymbol{X}^T\boldsymbol{a}+\boldsymbol{1}b\right)\right]\right\}\right\}.$$

Можно доказать, что здесь:

апостериорная ковариационная матрица $\boldsymbol{B}_{\boldsymbol{X},c}=\left(\frac{1}{c}\boldsymbol{X}\boldsymbol{X}^T+2\boldsymbol{I}\right)^{-1}$,

апостериорное условное математическое ожидание $\tilde{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\boldsymbol{a})=\hat{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}-\frac{1}{2Nc}\boldsymbol{1}^T\boldsymbol{X}^T(\boldsymbol{a}-\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c})$,

апостериорная условная дисперсия $\tilde{\sigma}^2_N=\frac{1}{2N}$.

Приближенная функция правдоподобия

Исходный критерий максимума правдоподобия:

$$(\hat{\mu},\hat{c}) = \operatorname{argmax} F(\boldsymbol{X},\boldsymbol{Y} \mid \mu, c) = \operatorname{argmax} \frac{\Phi(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{a}, b, c)\Psi(\boldsymbol{a}, b \mid \mu)}{P(\boldsymbol{a}, b \mid \boldsymbol{X},\boldsymbol{Y}, \mu, c)}.$$

Приближенная функция правдоподобия

Исходный критерий максимума правдоподобия:

$$(\hat{\mu},\hat{c}) = \operatorname{argmax} F(\boldsymbol{X},\boldsymbol{Y} \mid \mu, c) = \operatorname{argmax} \frac{\Phi(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{a}, b, c)\Psi(\boldsymbol{a}, b \mid \mu)}{P(\boldsymbol{a}, b \mid \boldsymbol{X},\boldsymbol{Y},\mu, c)}.$$

Приближенный критерий максимума правдоподобия:

$$(\hat{\mu},\hat{c}) = \operatorname{argmax} \hat{F}(\boldsymbol{X},\boldsymbol{Y} \mid \mu, c) = \operatorname{argmax} \frac{\Phi(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{a}, b, c)\Psi(\boldsymbol{a}, b \mid \mu)}{\mathcal{N}(\boldsymbol{a} \mid \hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c}, \boldsymbol{B}_{\boldsymbol{X},c})\mathcal{N}(b \mid \tilde{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\boldsymbol{a}), \tilde{\sigma}_N^2)}.$$

Приближенная функция правдоподобия

Исходный критерий максимума правдоподобия:

$$(\hat{\mu},\hat{c}) = \operatorname{argmax} F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c) = \operatorname{argmax} \frac{\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\Psi(\boldsymbol{a},b|\mu)}{P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)}.$$

Приближенный критерий максимума правдоподобия:

$$(\hat{\mu},\hat{c}) = \operatorname{argmax} \hat{F}(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c) = \operatorname{argmax} \frac{\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\Psi(\boldsymbol{a},b|\mu)}{\mathcal{N}(\boldsymbol{a}\,|\,\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c},\boldsymbol{B}_{\boldsymbol{X},c})\mathcal{N}(b\,|\,\tilde{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\boldsymbol{a}),\tilde{\sigma}_N^2)}.$$

EM-алгоритм для приближенного критерия максимума правдоподобия:

$$\hat{\mu}_{k+1} = \arg\max_{\mu} \int_{\mathbb{R}}\int_{\mathbb{R}^n} \left[\ln \Psi(\boldsymbol{a},b|\mu)\right] \mathcal{N}(\boldsymbol{a}\,|\,\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c},\boldsymbol{B}_{\boldsymbol{X},c})\mathcal{N}(b\,|\,\tilde{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\boldsymbol{a}),\tilde{\sigma}_N^2)\,d\boldsymbol{a}db,$$

$$\hat{c}_{k+1} = \arg\max_{c} \int_{\mathbb{R}}\int_{\mathbb{R}^n} \left[\ln \Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\right] \mathcal{N}(\boldsymbol{a}\,|\,\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c},\boldsymbol{B}_{\boldsymbol{X},c})\mathcal{N}(b\,|\,\tilde{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\boldsymbol{a}),\tilde{\sigma}_N^2)\,d\boldsymbol{a}db.$$

Приближенная функция правдоподобия

Исходный критерий максимума правдоподобия:

$$(\hat{\mu},\hat{c}) = \operatorname{argmax} F(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c) = \operatorname{argmax} \frac{\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\Psi(\boldsymbol{a},b|\mu)}{P(\boldsymbol{a},b\,|\,\boldsymbol{X},\boldsymbol{Y},\mu,c)}.$$

Приближенный критерий максимума правдоподобия:

$$(\hat{\mu},\hat{c}) = \operatorname{argmax} \hat{F}(\boldsymbol{X},\boldsymbol{Y}\,|\,\mu,c) = \operatorname{argmax} \frac{\Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\Psi(\boldsymbol{a},b|\mu)}{\mathcal{N}(\boldsymbol{a}\,|\,\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c},\boldsymbol{B}_{\boldsymbol{X},c})\mathcal{N}(b\,|\,\tilde{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\boldsymbol{a}),\tilde{\sigma}_N^2)}.$$

EM-алгоритм для приближенного критерия максимума правдоподобия:

$$\hat{\mu}_{k+1} = \arg\max_{\mu} \int_{\mathbb{R}}\int_{\mathbb{R}^n} \left[\ln \Psi(\boldsymbol{a},b|\mu)\right] \mathcal{N}(\boldsymbol{a}\,|\,\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c},\boldsymbol{B}_{\boldsymbol{X},c})\mathcal{N}(b\,|\,\tilde{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\boldsymbol{a}),\tilde{\sigma}_N^2)\,d\boldsymbol{a}db,$$

$$\hat{c}_{k+1} = \arg\max_{c} \int_{\mathbb{R}}\int_{\mathbb{R}^n} \left[\ln \Phi(\boldsymbol{X},\boldsymbol{Y}\,|\,\boldsymbol{a},b,c)\right] \mathcal{N}(\boldsymbol{a}\,|\,\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\mu,c},\boldsymbol{B}_{\boldsymbol{X},c})\mathcal{N}(b\,|\,\tilde{b}_{\boldsymbol{y},\boldsymbol{X},\mu,c}(\boldsymbol{a}),\tilde{\sigma}_N^2)\,d\boldsymbol{a}db.$$

Численная реализация такого алгоритма не составляет проблемы.

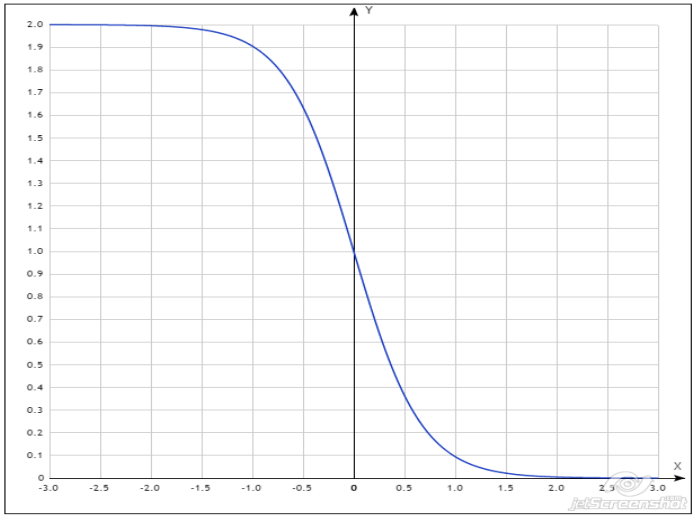
Приближенная функция правдоподобия

Шаг по параметру селективности

$$\hat{\mu}_{k+1} = \arg\min_{\mu \geq 0} \underbrace{\left[-\frac{1}{\sqrt{\pi}} \exp\left(-\frac{\mu^2}{4} \right) \left[\operatorname{erfc}\left(\frac{\mu}{2} \right) \right]^{-1} + h_k \mu \right]}_{\text{выпуклая функция}}$$

Здесь: $\operatorname{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^{\infty} \exp(-t^2) dt$ – функция ошибок,

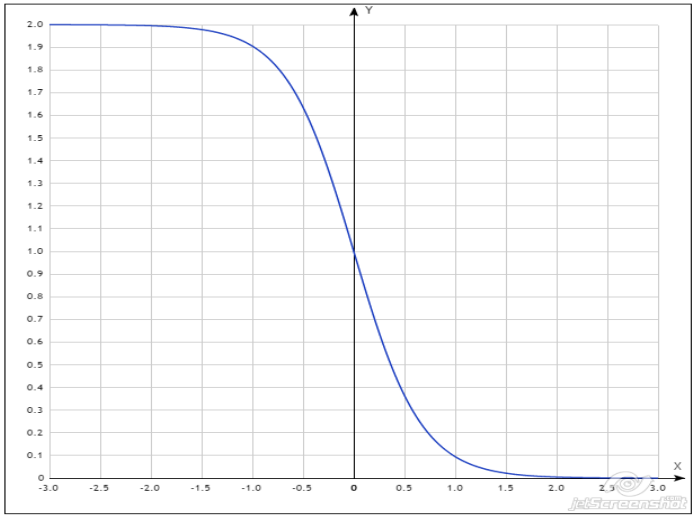
$h_k(\hat{a}_{i,y,\boldsymbol{x},\hat{\mu}_k,\hat{c}_k}, \tilde{\sigma}_N^2)$ – константа, определяемая аппроксимирующим нормальным распределением.



Приближенная функция правдоподобия

Шаг по параметру селективности

$$\hat{\mu}_{k+1} = \arg\min_{\mu \geq 0} \underbrace{\left[-\frac{1}{\sqrt{\pi}} \exp\left(-\frac{\mu^2}{4} \right) \left[\operatorname{erfc}\left(\frac{\mu}{2} \right) \right]^{-1} + h_k \mu \right]}_{\text{выпуклая функция}}$$



Здесь: $\operatorname{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^{\infty} \exp(-t^2) dt$ – функция ошибок,

$h_k(\hat{a}_{i,\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k}, \tilde{\sigma}_N^2)$ – константа, определяемая аппроксимирующим нормальным распределением.

Шаг по параметру наблюдений

Обучающая совокупность $\left\{ (\boldsymbol{x}_j, y_j),\ j = 1,\ldots,N \right\}, \quad \boldsymbol{x}_j \in \mathbb{R}^n,\ y_j \in \mathbb{R}$

$$\hat{c}_{k+1} = \frac{1}{N} \left(\boldsymbol{y} - (\boldsymbol{X}^T \hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k} + \boldsymbol{1}\hat{b}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k}) \right)^T \left(\boldsymbol{y} - (\boldsymbol{X}^T \hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k} + \boldsymbol{1}\hat{b}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k}) \right) +$$

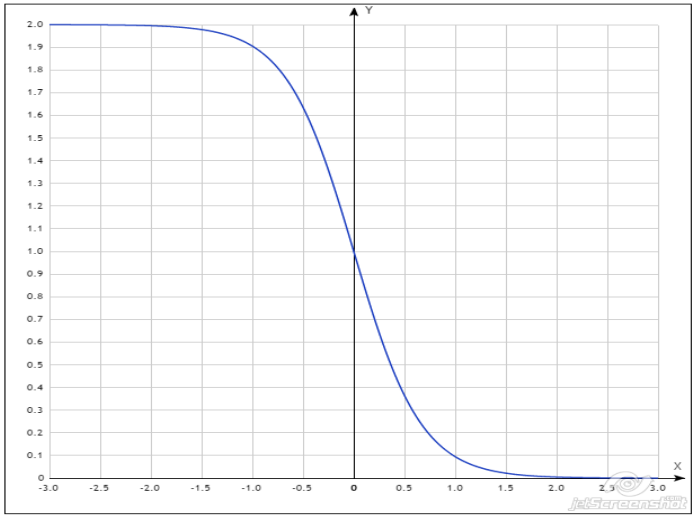
$$\frac{1}{4\hat{c}_k} Tr\left[\left(\frac{1}{N^2} \sum_{j=1}^{N} \sum_{l=1}^{N} \boldsymbol{x}_j \boldsymbol{x}_l^T \right) \left(\boldsymbol{X}\boldsymbol{X}^T + 2\hat{c}_k \boldsymbol{I} \right)^{-1} \right] + \tilde{\sigma}_N^2 .$$

Здесь $\left(\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k}, \hat{b}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k} \right)$ – результат обучения на предыдущем шаге.

Приближенная функция правдоподобия

Шаг по параметру селективности

$$\hat{\mu}_{k+1} = \arg\min_{\mu \geq 0} \underbrace{\left[-\frac{1}{\sqrt{\pi}} \exp\left(-\frac{\mu^2}{4} \right) \left[\operatorname{erfc}\left(\frac{\mu}{2} \right) \right]^{-1} + h_k \mu \right]}_{\text{выпуклая функция}}$$



Здесь: $\operatorname{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^{\infty} \exp(-t^2) dt$ – функция ошибок,

$h_k(\hat{a}_{i,\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k}, \tilde{\sigma}_N^2)$ – константа, определяемая аппроксимирующим нормальным распределением.

Шаг по параметру наблюдений

Обучающая совокупность $\{(\boldsymbol{x}_j, y_j),\ j = 1,\ldots,N\}, \quad \boldsymbol{x}_j \in \mathbb{R}^n,\ y_j \in \mathbb{R}$

$$\hat{c}_{k+1} = \frac{1}{N}\left(\boldsymbol{y} - (\boldsymbol{X}^T \hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k} + \boldsymbol{1}\hat{b}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k}) \right)^T \left(\boldsymbol{y} - (\boldsymbol{X}^T \hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k} + \boldsymbol{1}\hat{b}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k}) \right) +$$

$$\frac{1}{4\hat{c}_k} Tr\left[\left(\frac{1}{N^2} \sum_{j=1}^{N} \sum_{l=1}^{N} \boldsymbol{x}_j \boldsymbol{x}_l^T \right) \left(\boldsymbol{X}\boldsymbol{X}^T + 2\hat{c}_k \boldsymbol{I} \right)^{-1} \right] + \tilde{\sigma}_N^2 .$$

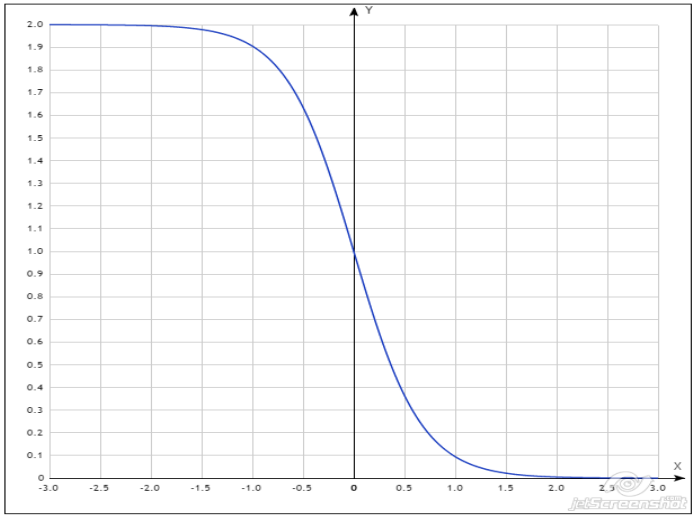
Здесь $\left(\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k}, \hat{b}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k} \right)$ – результат обучения на предыдущем шаге.

Алгоритм не требует перебора значений структурных параметров (μ, c), как в известном методе Regularization Path.

Приближенная функция правдоподобия

Шаг по параметру селективности

$$\hat{\mu}_{k+1} = \arg\min_{\mu \geq 0} \underbrace{\left[-\frac{1}{\sqrt{\pi}} \exp\left(-\frac{\mu^2}{4} \right) \left[\operatorname{erfc}\left(\frac{\mu}{2} \right) \right]^{-1} + h_k \mu \right]}_{\text{выпуклая функция}}$$



Здесь: $\operatorname{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^{\infty} \exp(-t^2) dt$ – функция ошибок,

$h_k(\hat{a}_{i,\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k}, \tilde{\sigma}_N^2)$ – константа, определяемая аппроксимирующим нормальным распределением.

Шаг по параметру наблюдений

Обучающая совокупность $\{(\boldsymbol{x}_j, y_j),\ j = 1,\ldots,N\}, \quad \boldsymbol{x}_j \in \mathbb{R}^n,\ y_j \in \mathbb{R}$

$$\hat{c}_{k+1} = \frac{1}{N}\left(\boldsymbol{y} - (\boldsymbol{X}^T \hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k} + \boldsymbol{1}\hat{b}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k}) \right)^T \left(\boldsymbol{y} - (\boldsymbol{X}^T \hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k} + \boldsymbol{1}\hat{b}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k}) \right) +$$

$$\frac{1}{4\hat{c}_k} Tr\left[\left(\frac{1}{N^2} \sum_{j=1}^{N} \sum_{l=1}^{N} \boldsymbol{x}_j \boldsymbol{x}_l^T \right) \left(\boldsymbol{X}\boldsymbol{X}^T + 2\hat{c}_k \boldsymbol{I} \right)^{-1} \right] + \tilde{\sigma}_N^2 .$$

Здесь $\left(\hat{\boldsymbol{a}}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k}, \hat{b}_{\boldsymbol{y},\boldsymbol{X},\hat{\mu}_k,\hat{c}_k} \right)$ – результат обучения на предыдущем шаге.

J. Friedman, T. Hastie, R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 2010, Volume 33, Issue 1.