Kyoto University

統計的モデリング基礎⑧
～正則化と事後確率最大化～

鹿島久嗣
（情報学科 計算機科学コース）

Department of Intelligence Science
and Technology

正則化

重回帰モデルの復習：最小二乗法による定式化

- 重回帰モデル：$y = \boldsymbol{\beta}^\top \mathbf{x}$
  - パラメータ： $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_m, \alpha)^\top$
  - 独立変数：$\mathbf{x} = (x_1, x_2, \dots, x_m, 1)^\top$

  最後の次元は切片部分に相当
- データ：
  - 計画行列： $\boldsymbol{X} = \left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}\right)^\top$
  - 従属変数： $\mathbf{y} = \left(y^{(1)}, y^{(2)}, \dots, y^{(n)}\right)^\top$
- 目的関数：$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\beta}^\top \mathbf{x}^{(i)}\right)^2 = \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$

重回帰モデルの解：
解析解が得られる

- 目的関数：$L(\boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})$
- 解: $\boldsymbol{\beta}^* = \mathrm{argmin}_{\boldsymbol{\beta}}\, L(\boldsymbol{\beta}) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}$
- ただし、解が存在するためには$\boldsymbol{X}^\top\boldsymbol{X}$が正則である必要
  - モデルの次元数mよりもデータ数nが大きい場合はおおむね成立
- 正則化（regularization）：正則でない場合には$\boldsymbol{X}^\top\boldsymbol{X}$の対角成分に正の定数$\lambda > 0$を加えて正則にする
  - 新たな解：$\boldsymbol{\beta}^* = \mathrm{argmin}_{\boldsymbol{\beta}}\, L(\boldsymbol{\beta}) = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\mathbf{y}$
  - 目的関数に戻ると：$L(\boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$

パラメータのノルムに関するペナルティ項

データへの過適合：
データへの過剰な適合は将来のデータへの予測力を損なう

- 先ほどは、正則化を計算を安定させるために導入した
  - 例えば、データ数nが次元数mより小さいとき、重回帰の解は一意に定まらない
  - 任意の数の解が存在し、どれが良いのかわからない
- データへの過適合：
  - 汎化（generalization）：予測を目的とする場合、我々の真の目的は将来のデータへの予測力が高いモデルを得ること
  - 手持ちのデータへのモデルの過剰な適合は、将来の予測力を損なう可能性がある

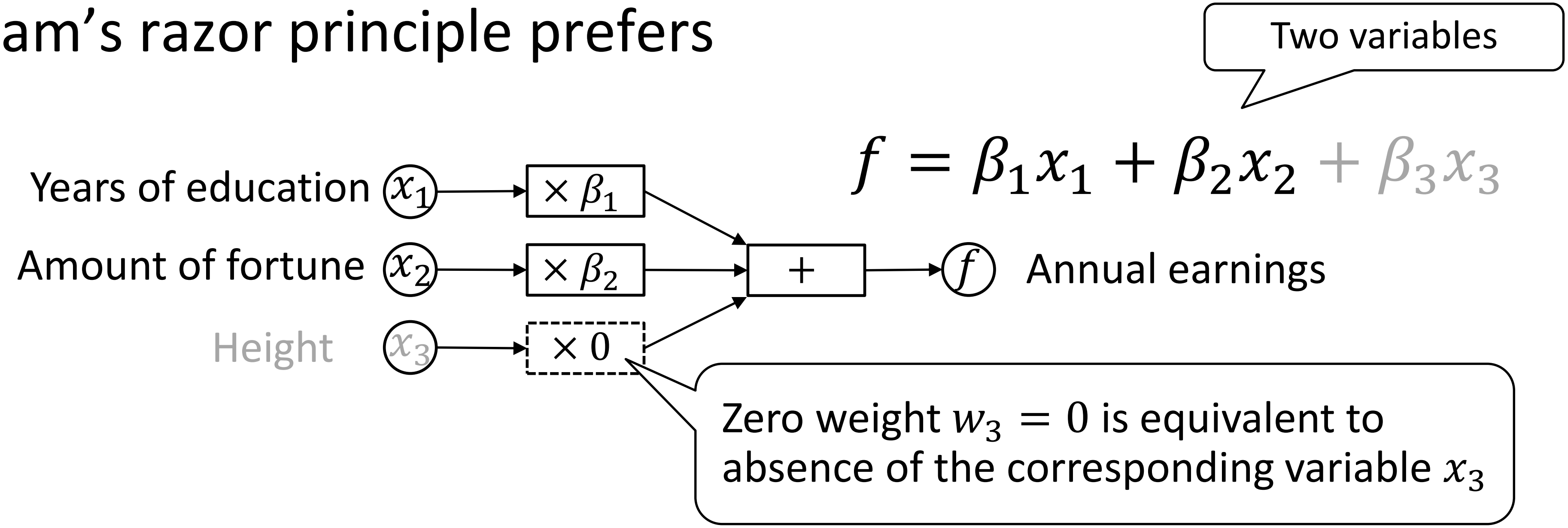
オッカムの剃刀：
できるだけ“単純な“なモデルを採用せよ

オッカムのウィリアム

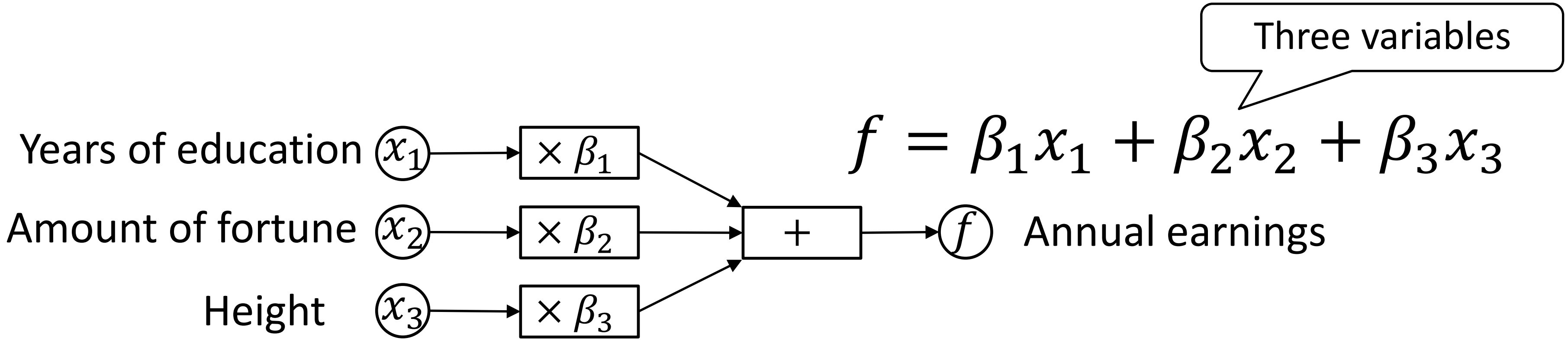
- データに同程度適合している無数のモデルのうち、どれが最も“良い” モデルだろうか
- オッカムの剃刀：単純なモデルを採用せよ
  - 「ある事柄を説明するために必要以上に多くを仮定すべきでない」
- 単純さを何で測るか？
  - 例えば、モデルに含まれる独立変数の数
  - 自由度調整済決定係数、AICやBICなどの情報量基準を用いる

Occam's razor principle:
Prefers models with smaller number of variables

- Occam's razor principle prefers



to



0-ノルムを用いた正則化：パラメータ中の非零成分の数を減らす

- モデルの単純さの指標： モデルに含まれる独立変数の数
  = $\boldsymbol{\beta}$ 中の非零成分の数 = $\boldsymbol{\beta}$ の0-ノルム
- 0-ノルム制約を入れた回帰問題：

モデルに含める独立変数の数

$$\boldsymbol{\beta}^* = \mathrm{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \ \ \mathrm{s.t.}\ \|\boldsymbol{\beta}\|_0 \le \eta$$

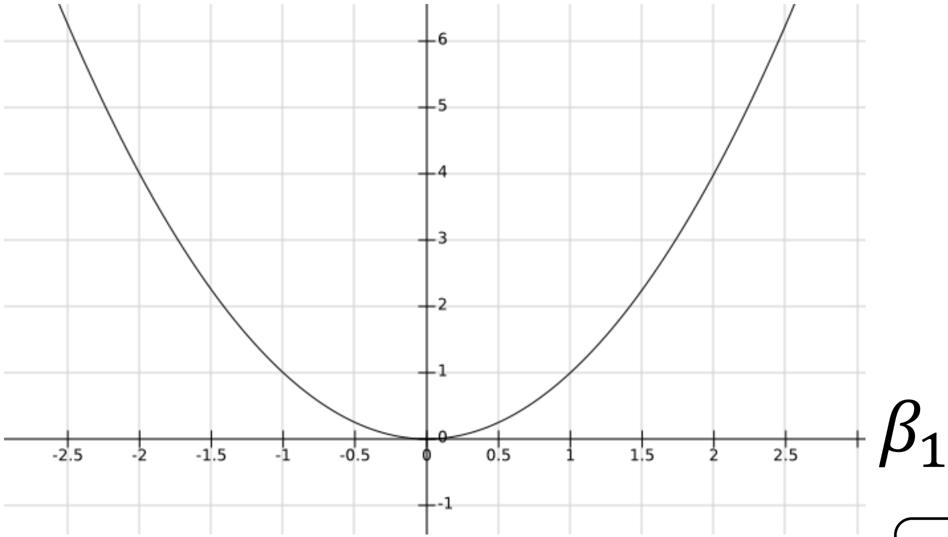
- あるいは 0-ノルムをペナルティ項として導入：

$$\boldsymbol{\beta}^* = \mathrm{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_0$$

  - η と λ は一対一対応している
- ただし、この問題は凸最適化問題でないため、扱いが大変

0-ノルムの代替：
2-ノルム正則化は凸最適化になる

- 0-ノルム $\|\boldsymbol{\beta}\|_0$の代わりに 2-ノルム $\|\boldsymbol{\beta}\|_2^2$ を用いる



- リッジ回帰：$L(\boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$ 凸 ☺
  - 0-ノルム正則化の緩和版として捉えることができる：

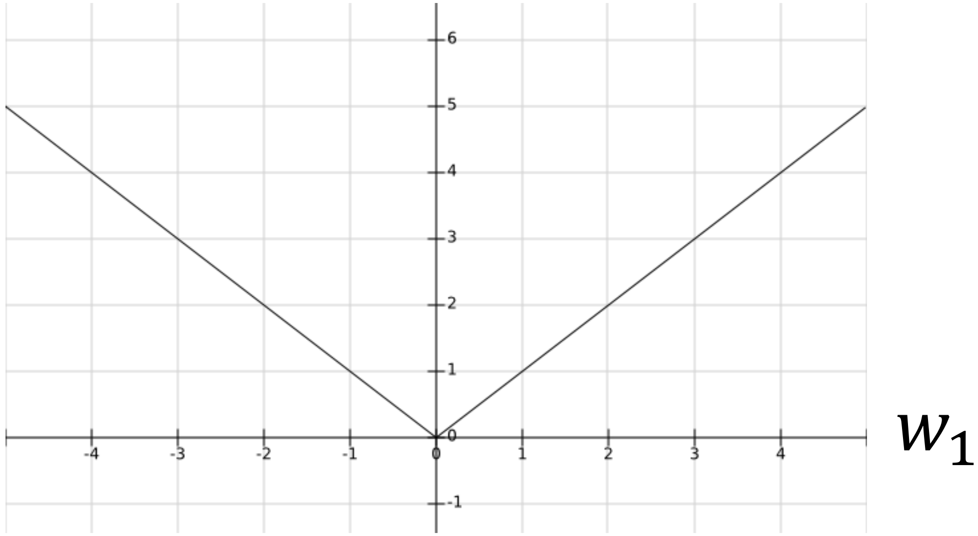
$$L(\boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_0$$

非凸 ☹

 KYOTO UNIVERSITY

0-ノルムの代替：1-ノルム正則化も凸最適化になる

- さらに、1-ノルム $\|\boldsymbol{\beta}\|_1 = |\beta_1| + |\beta_2| + \cdots + |\beta_D|$ も利用可能



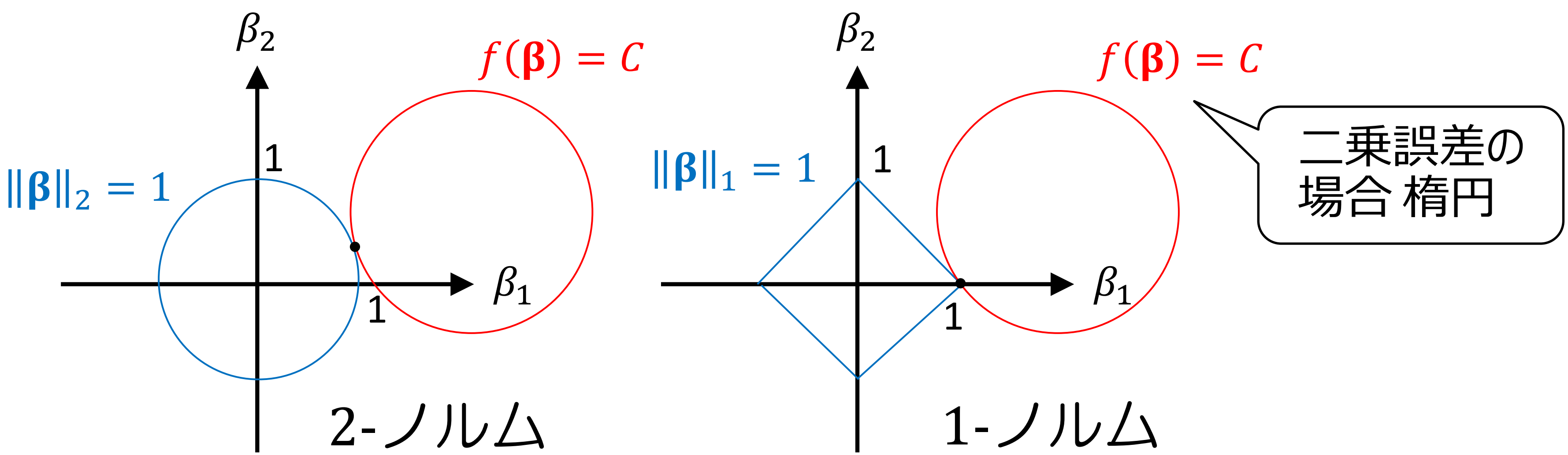
- ラッソ：$L(\boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$
  - 凸最適化だが、解析解はもたない
- 1-ノルムを用いると疎な解になる（$\boldsymbol{\beta}^*$の多くの要素が0になる）

1-ノルム正則化の利点：疎な解をもつ

- 1-ノルム正則化は1-ノルム制約で書き直せる：

$$\mathrm{argmin}_{\boldsymbol{\beta}}\, f(\boldsymbol{\beta}) + \gamma \|\boldsymbol{\beta}\|_1 \Leftrightarrow \mathrm{argmin}_{\boldsymbol{\beta}}\, f(\boldsymbol{\beta}) \text{ s.t. } \|\boldsymbol{\beta}\|_1 \le \lambda$$

- 1-ノルム正則化は疎な最適解をもつ傾向がある
  - 2-ノルム制約の等高線（円形）と1-ノルム制約の等高線（菱形）の比較



事後確率最大化推定

回帰のベイズ統計的解釈：事後確率最大化推定

- 線形回帰モデルの推定は最尤推定として解釈できた
- 正則化のもとでの回帰モデルの推定はベイズ統計の枠組みで解釈できる
  - 事前分布・事後分布の導入
  - 事後確率最大化（MAP）推定
  - リッジ回帰＝MAP推定

線形回帰モデルの最尤推定：
線形回帰の確率モデルの最尤推定 ＝ 最小二乗法

- 線形回帰モデルに対応する確率モデル
  - 確率密度関数：$f\left(y^{(i)} \mid \mathbf{x}^{(i)}\right) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\left(y^{(i)} - \boldsymbol{\beta}^\top \mathbf{x}^{(i)}\right)^2}{2\sigma^2}\right)$
- 対数尤度：$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f\left(y^{(i)} \mid \mathbf{x}^{(i)}\right)$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\beta}^\top \mathbf{x}^{(i)}\right)^2 + \mathrm{const.}$$

- 対数尤度を$\boldsymbol{\beta}$について最大化すること（最尤推定）
  ＝二乗誤差を$\boldsymbol{\beta}$について最小化すること（最小二乗法）

ベイズ的統計モデリングの考え方：
尤度の代わりに事後分布を考える

- 最尤推定（MLE）では尤度を最大化するパラメータ$\boldsymbol{\beta}$を求めた：

$$P(\mathbf{y} \mid \boldsymbol{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} f(y^{(i)} | \mathbf{x}^{(i)}, \boldsymbol{\beta})$$

  - これは、パラメータが与えられたもとでデータが再現される条件付確率: $P\left(\text{データ} \mid \text{パラメータ}\right) = P(\mathbf{y} \mid \boldsymbol{X}, \boldsymbol{\beta})$

    ※ ここでは、$\boldsymbol{X}$は定数として与えられ、$\mathbf{y}$のみをデータ（確率変数）として扱っている点に注意

- ベイズ的な統計モデリングの考え方では、事後分布を考える：

$$P\left(\text{パラメータ} \mid \text{データ}\right) = P(\boldsymbol{\beta} \mid \boldsymbol{X}, \mathbf{y})$$

  - 事後分布はパラメータを確率変数と考える

事後分布：
事後分布 ∝ 尤度 × 事前分布

- 事後分布をベイズの定理で書き換えると：

$$P(\text{パラメータ} \mid \text{データ}) = \frac{P(\text{データ} \mid \text{パラメータ})P(\text{パラメータ})}{P(\text{データ})}$$

ベイズの定理

パラメータに依存しない
（定数扱い）

  - $P(\text{データ}) = \sum_{\text{パラメータ}} P(\text{データ} \mid \text{パラメータ})P(\text{パラメータ})$

- 対数事後分布：

$$\log P(\text{パラメータ} \mid \text{データ})$$
$$= \underbrace{\log P(\text{データ} \mid \text{パラメータ})}_{\text{尤度}} + \underbrace{\log P(\text{パラメータ})}_{\text{事前分布}} + \text{const.}$$

事後確率最大化（MAP）推定：事後確率を最大化するパラメータを採用

- 対数事後分布：
  $$\log P(\text{パラメータ} \mid \text{データ})$$
  $$= \log P(\text{データ} \mid \text{パラメータ}) + \log P(\text{パラメータ}) + \text{const.}$$
- 事後確率最大化（Maximum a posteriori; MAP）推定
  - 事後確率を最大化するパラメータを採用する：
    $$\text{パラメータ}^* = \operatorname{argmax}_{\text{パラメータ}} \log P(\text{パラメータ} \mid \text{データ})$$
  - 最尤推定では $\log P(\text{データ} \mid \text{パラメータ})$ の項のみを考える
  - 追加の項として事前分布の項：$\log P(\text{パラメータ})$

事後確率最大化としてのリッジ回帰：正規分布を事前分布とした事後確率最大化

- 対数事後分布：

$$\log P(\text{パラメータ} \mid \text{データ})$$
$$= \underbrace{\log P(\text{データ} \mid \text{パラメータ})}_{\text{対数尤度}} + \underbrace{\log P(\text{パラメータ})}_{\text{対数事前分布}} + \text{const.}$$

- リッジ回帰：

$$\boldsymbol{\beta}^* = \operatorname{argmin}_{\boldsymbol{\beta}} \frac{1}{2{\sigma'}^2} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\beta}^\top \mathbf{x}^{(i)}\right)^2 + \frac{1}{2\sigma^2} \|\boldsymbol{\beta}\|_2^2$$

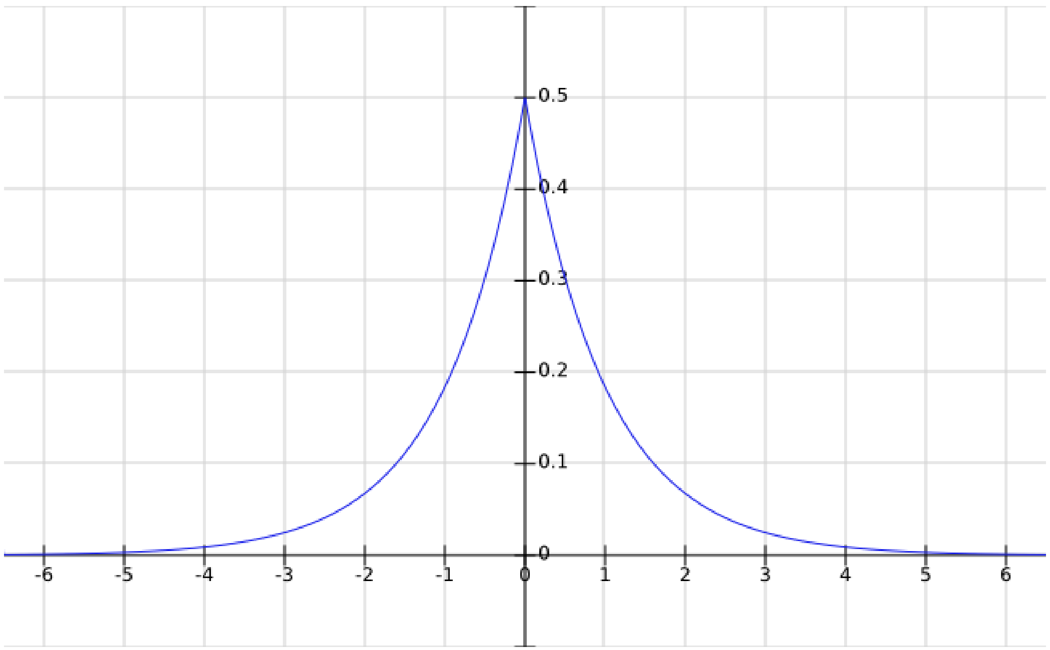
  - 対数尤度： $\sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma'} \exp\left(-\frac{\left(y^{(i)} - \boldsymbol{\beta}^\top \mathbf{x}^{(i)}\right)^2}{2{\sigma'}^2}\right)$
  - 事前分布： $P(\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\boldsymbol{\beta}^\top \boldsymbol{\beta}}{2\sigma^2}\right)$

平均**0**の
正規分布にとる

事後確率最大化としてのラッソ：ラプラス分布を事前分布として利用

- 事前分布を正規分布にすると2-ノルム正則化
- ラプラス分布：1-ノルム正則化に対応する事前分布

$$P(\boldsymbol{\beta}) = \frac{1}{2\phi}\exp\left(-\frac{|\boldsymbol{\beta}|}{\phi}\right)$$



ベイズ予測：推定のばらつきを考慮した予測

- MAP推定では事後分布が最大となるパラメータを点推定する

$$\hat{\boldsymbol{\beta}} = \mathrm{argmax}_{\boldsymbol{\beta}}\, P(\boldsymbol{\beta} \mid \boldsymbol{X}, \mathbf{y})$$

- ベイズ予測では事後分布そのものを用いて予測する

$$P(y \mid \mathbf{x}) = \int_{\boldsymbol{\beta}} P(y \mid \mathbf{x}, \boldsymbol{\beta})\, P(\boldsymbol{\beta} \mid \boldsymbol{X}, \mathbf{y})\, \mathrm{d}\boldsymbol{\beta}$$

  - あらゆるパラメータにおけるモデルの予測を事後確率で重みづけて予測する
  - 最適化問題を解いてパラメータを点推定するのでなく「全部」使う

まとめ：
正則化と事後確率最大化

- 正則化：データへの過適合を防ぎ、汎化を促進する
  - オッカムの剃刀：できるだけ単純なモデルを採用せよ
  - 0-ノルム正則化：含まれる独立変数の数は最適化困難
  - 2-ノルム正則化： 0-ノルムの凸緩和で扱いやすい → リッジ回帰
  - 1-ノルム正則化：凸かつ疎な解を得る効果あり → ラッソ
- 事後確率最大化（MAP）推定：
  - ベイズ統計では、パラメータの事後分布を考える
  - 事後確率最大化：事後確率を最大化するパラメータを良しとする
  - リッジ回帰・ラッソは事後確率最大化としても解釈できる