Linear Regression

Václav Šmídl

March 13, 2018

Linear regression and OLS

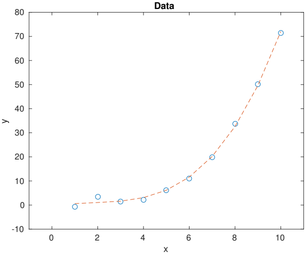
Fit by a linear function:

$$
\begin{array}{llll}
y_1 & = a x_1 & + b 1, & + e_1 \\
y_2 & = a x_2 & + b 1 & + e_2, \\
\vdots & \vdots & \vdots & \vdots
\end{array}
$$

In matrix notation $\theta = [a, b]^T$:

$$\mathbf{y} = X\theta + \mathbf{e},$$

Minimize $\sum_i e_i^2 = \mathbf{e}^T\mathbf{e} = ||\mathbf{y} - X\theta||_2^2$:



Linear regression and OLS

Fit by a linear function:

$$\begin{array}{llll} y_1 & = ax_1 & +b1, & +e_1 \\ y_2 & = ax_2 & +b1 & +e_2, \\ \vdots & \vdots & \vdots & \vdots \end{array}$$

In matrix notation $\theta = [a, b]^T$:

$$\mathbf{y} = X\theta + \mathbf{e},$$

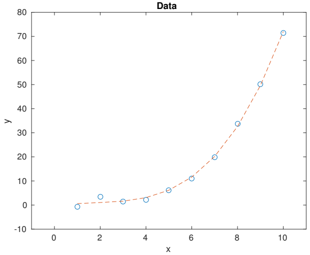
Minimize $\sum_i e_i^2 = \mathbf{e}^T\mathbf{e} = ||\mathbf{y} - X\theta||_2^2$:

$$\frac{d(\mathbf{e}^T\mathbf{e})}{d\theta} = 0.$$

$$\frac{d}{d\theta}((\mathbf{y} - X\theta)^T(\mathbf{y} - X\theta)) = 0$$

$$\frac{d}{d\theta}(\mathbf{y}^T\mathbf{y} - \theta^T X^T\mathbf{y} - \mathbf{y}^T X\theta + \theta^T X^T X\theta) = 0$$

$$-X^T\mathbf{y} + X^T X\theta = 0$$



Linear regression and OLS

Fit by a linear function:

$$\begin{aligned} y_1 &= a x_1 &+ b 1, &\quad + e_1 \\ y_2 &= a x_2 &+ b 1 &\quad + e_2, \\ \vdots & \quad \vdots & \vdots & \quad \vdots \end{aligned}$$

In matrix notation $\theta = [a, b]^T$:

$$\mathbf{y} = X\theta + \mathbf{e},$$

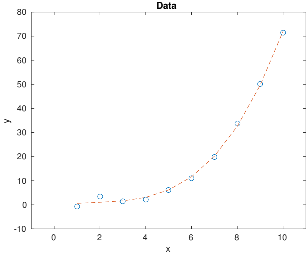
Minimize $\sum_i e_i^2 = \mathbf{e}^T\mathbf{e} = ||\mathbf{y} - X\theta||_2^2$:

$$\frac{d(\mathbf{e}^T\mathbf{e})}{d\theta} = 0.$$

$$\frac{d}{d\theta}((\mathbf{y} - X\theta)^T(\mathbf{y} - X\theta)) = 0$$

$$\frac{d}{d\theta}(\mathbf{y}^T\mathbf{y} - \theta^T X^T\mathbf{y} - \mathbf{y}^T X\theta + \theta^T X^T X\theta) = 0$$

$$-X^T\mathbf{y} + X^T X\theta = 0$$



Solution:

$$\hat{\theta} = (X^T X)^{-1} X^T \mathbf{y}.$$

Ridge regression

Ordinary least squares:

$$\hat{\theta} = (X^T X)^{-1} X^T \mathbf{y}.$$

problematic for numerical stability. For $\min \operatorname{eig}(X^T X) \to 0$,

Ridge regression

Ordinary least squares:

$$\hat{\theta} = (X^T X)^{-1} X^T \mathbf{y}.$$

problematic for numerical stability. For $\min \operatorname{eig}(X^T X) \to 0$, $\hat{\theta} \to \infty$.

Ridge regression

Ordinary least squares:

$$\hat{\theta} = (X^T X)^{-1} X^T \mathbf{y}.$$

problematic for numerical stability. For $\min \operatorname{eig}(X^T X) \to 0$, $\hat{\theta} \to \infty$.

- Replace inverse by pseudo-inverse (threshold),

Ridge regression

Ordinary least squares:

$$\hat{\theta} = (X^T X)^{-1} X^T \mathbf{y}.$$

problematic for numerical stability. For $\min \mathrm{eig}(X^T X) \to 0$, $\hat{\theta} \to \infty$.

- Replace inverse by pseudo-inverse (threshold),
- Add penalization for large values :

  $$\hat{\theta} = \arg\min_{\theta} \left(||\mathbf{y} - X\theta||_2^2 + \alpha ||\theta||_2^2 \right),$$

  where α is a suitably chosen coefficient.
- Solution

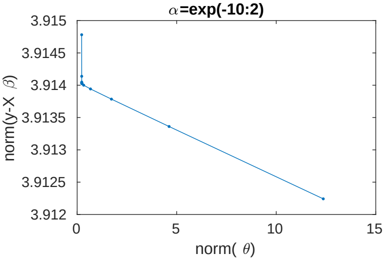
  $$\hat{\theta} = (X^T X + \alpha I)^{-1} X^T \mathbf{y}.$$

  minimal eigenvalue is α.

Ridge regression – selection of α

Selection of α:

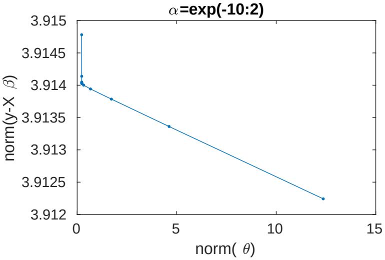
1. Cross-validation
   (analytical solution for α)
2. L-curve
   plot of $||\theta||_2^2$ versus $||\mathbf{y} - X\theta||_2^2$
3. Bayes



Ridge regression – selection of α

Selection of α:

1. Cross-validation
   (analytical solution for α)
2. L-curve
   plot of $||\theta||_2^2$ versus $||\mathbf{y} - X\theta||_2^2$
3. Bayes



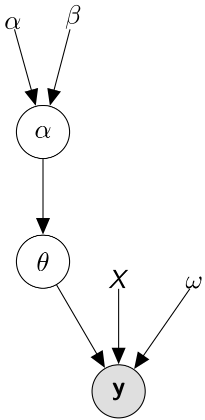
Probability model

$$
\begin{aligned}
p(\mathbf{y}, \theta | X, \alpha) &= p(\mathbf{y}|\theta, X) p(\theta|\alpha) \\
&= \mathcal{N}(X\theta, I)\mathcal{N}(0, \alpha^{-1} I) \\
&\propto \exp\left\{ -\frac{1}{2}||\mathbf{y} - X\theta||_2^2 - \frac{1}{2}\alpha||\theta||_2^2 \right\}
\end{aligned}
$$

Ridge regression – selection of α

Probability model

$$\begin{aligned} p(\mathbf{y}, \theta | X, \alpha) &= p(\mathbf{y}|\theta, X) p(\theta|\alpha) \\ &= \mathcal{N}(X\theta, I)\mathcal{N}(0, \alpha^{-1} I) \\ &\propto \exp\left\{ -\frac{1}{2} ||\mathbf{y} - X\theta||_2^2 - \frac{1}{2}\alpha ||\theta||_2^2 \right\} \end{aligned}$$



Introduce prior

$$p(\alpha) = G(\delta, \gamma),$$

compute

$$p(\alpha|\mathbf{y}, X)$$

or

$$p(\theta|X, \mathbf{y}) = \int p(\theta, \alpha|X, \mathbf{y}) d\alpha$$

.

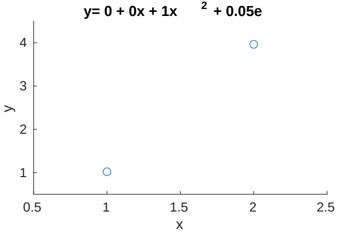
Ridge regression – polynomial case

Define a degenerate case:

- true model

$$y = x^2 + e$$

- observations at $x = [1, 2]$.
- fit polynomial of 5th order.
- Model selection?



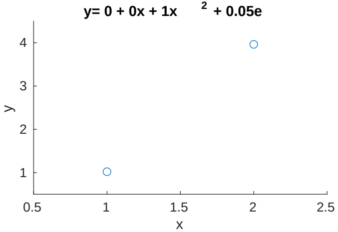
Ridge regression – polynomial case

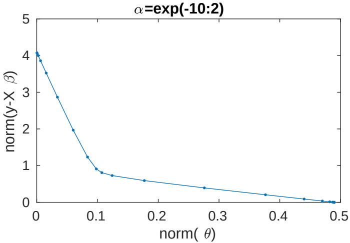
Define a degenerate case:

- true model

$$y = x^2 + e$$

- observations at $x = [1, 2]$.
- fit polynomial of 5th order.
- Model selection?





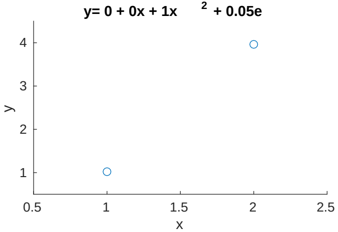
Ridge regression – polynomial case

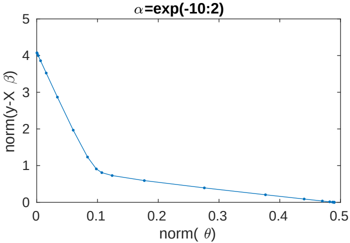
Define a degenerate case:

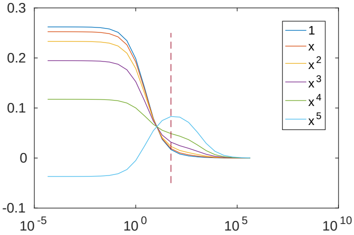
- true model

$$y = x^2 + e$$

- observations at $x = [1, 2]$.
- fit polynomial of 5th order.
- Model selection?



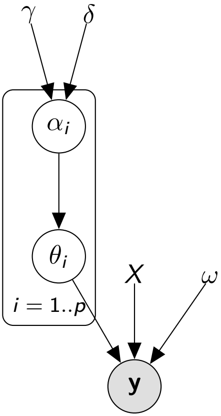




Automatic relevance determination (ARD)

Probability model

$$
\begin{aligned}
p(\mathbf{y}, \theta | X, \alpha) &= p(\mathbf{y}|\theta, X) p(\theta|\alpha) \\
&= \mathcal{N}(X\theta, I)\mathcal{N}(0, \mathrm{diag}[\alpha_1, \ldots, \alpha_p]) \\
&\propto \exp\left\{ -\frac{1}{2}||\mathbf{y} - X\theta||_2^2 - \frac{1}{2}\sum_i \alpha_i \theta_i^2 \right\}
\end{aligned}
$$



Introduce prior

$$p(\alpha_i) = G(\delta, \gamma), \;\; p(\boldsymbol{\alpha}) = \prod_i p(\alpha_i)$$

compute

$$p(\boldsymbol{\alpha}|\mathbf{y}, X)$$

or

$$p(\theta|X, \mathbf{y}) = \int p(\theta, \alpha | X, \mathbf{y}) d\alpha$$

.

Variational Bayes for Automatic relevance determination

Probability model

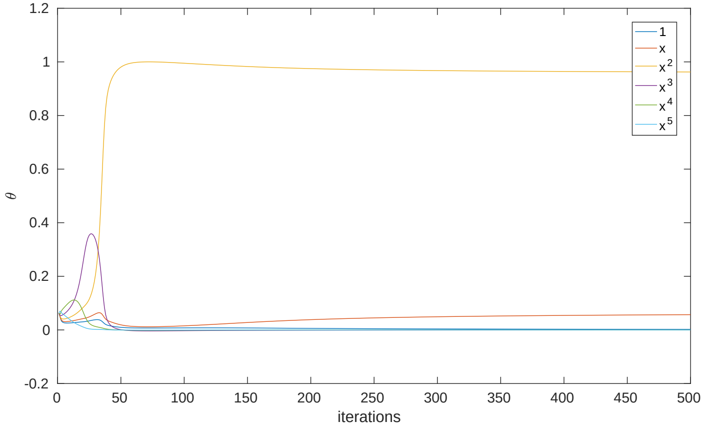
$$p(\mathbf{y}, \theta | X, \alpha) = \mathcal{N}(X\theta, I)\mathcal{N}(0, \text{diag}[\alpha_1, \ldots, \alpha_p]) \prod_i G(\delta, \gamma)$$

Posterior factors

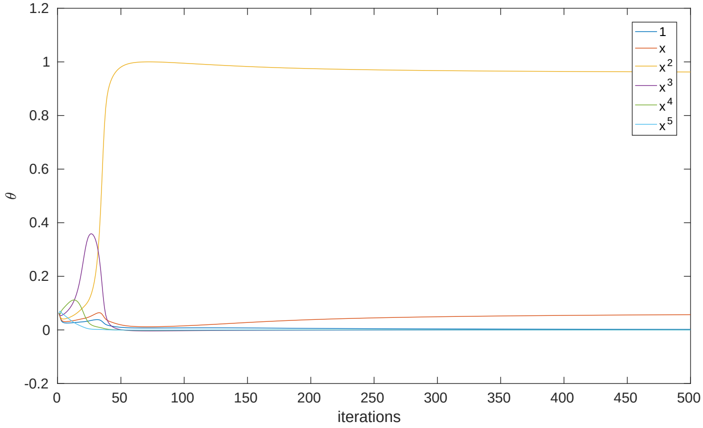
$$\begin{aligned} p(\alpha_i | \mathbf{y}, X) &= G(\delta, \gamma_i), \\ \gamma_i &= \gamma_0 + \frac{1}{2} \langle \theta_i^2 \rangle \\ p(\theta | \mathbf{y}, X) &= \mathcal{N}(\hat{\theta}, \Sigma_\theta), \\ \hat{\theta} &= (X^T X + \text{diag} \langle \alpha \rangle)^{-1} X^T \mathbf{y}, \\ \Sigma_\theta &= (X^T X + \text{diag} \langle \alpha \rangle)^{-1}. \end{aligned}$$

Iterated least squares.

Variational Bayes for Automatic relevance determination

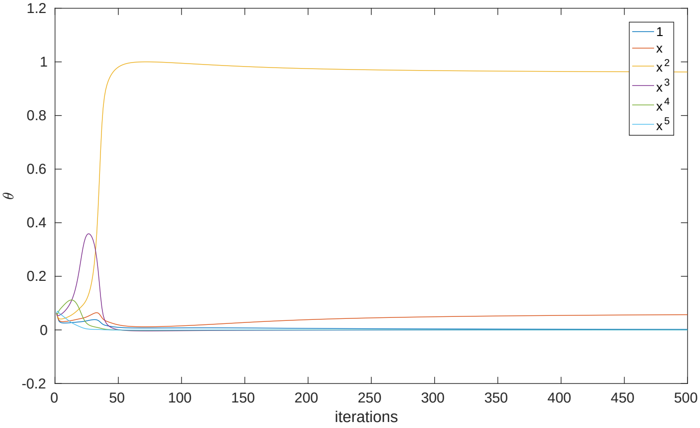


Variational Bayes for Automatic relevance determination



- What is penalized?

Variational Bayes for Automatic relevance determination



- What is penalized?
- Is it better in the sense of marginal likelihood (BIC) then original model?

Revisiting toy example:

Noisy observation:

$$y = 1\theta + e,$$

where:

$$\begin{aligned} p(y) &= \mathcal{N}(\theta, 1), \\ p(\theta|\alpha) &= \mathcal{N}(0, \alpha^{-1}), \\ p(\alpha) &= G(0_+, 0_+). \end{aligned}$$

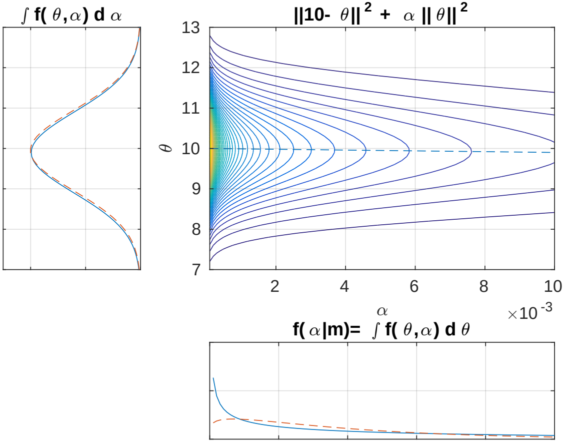
Revisiting toy example:

Noisy observation:

$$y = 1\theta + e,$$

where:

$$p(y) = \mathcal{N}(\theta, {\color{red}1}),$$
$$p(\theta|\alpha) = \mathcal{N}(0, \alpha^{-1}),$$
$$p(\alpha) = G(0_+, 0_+).$$



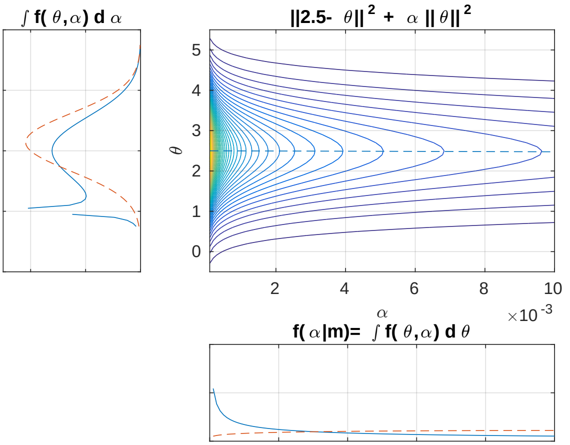
Revisiting toy example:

Noisy observation:

$$y = 1\theta + e,$$

where:

$$p(y) = \mathcal{N}(\theta, {\color{red}1}),$$
$$p(\theta|\alpha) = \mathcal{N}(0, \alpha^{-1}),$$
$$p(\alpha) = G(0_+, 0_+).$$



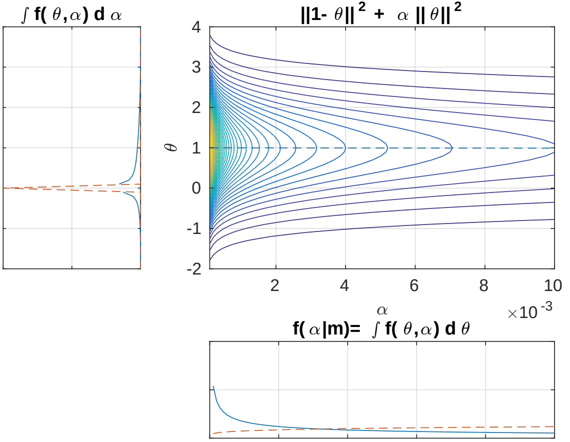
Revisiting toy example:

Noisy observation:

$$y = 1\theta + e,$$

where:

$$p(y) = \mathcal{N}(\theta, {\color{red}1}),$$
$$p(\theta|\alpha) = \mathcal{N}(0, \alpha^{-1}),$$
$$p(\alpha) = G(0_+, 0_+).$$



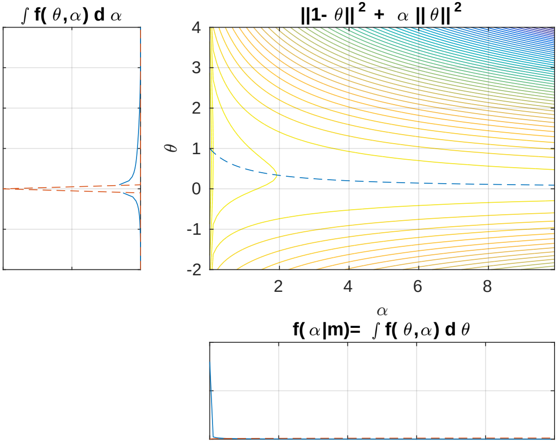
Revisiting toy example:

Noisy observation:

$$y = 1\theta + e,$$

where:

$$p(y) = \mathcal{N}(\theta, \color{red}{1}),$$
$$p(\theta|\alpha) = \mathcal{N}(0, \alpha^{-1}),$$
$$p(\alpha) = G(0_+, 0_+).$$



Sparse Linear Regression:

Prior has peak at zero and heavy tail

ARD prior:

$$p(\theta) = \int p(\theta|\alpha)p(\alpha)d\alpha = St(0, \sigma, \nu),$$

with two possible hidden variable formulations.

Laplace prior

$$p(\theta) = (2b)^{-1} \exp\left(-\frac{1}{2b}|x|\right).$$

with joint likelihood (LASSO)

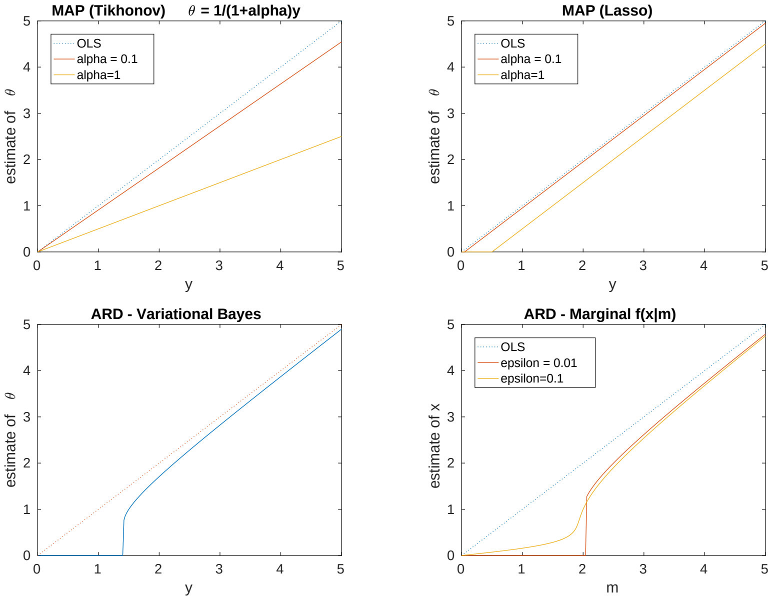
$$p(\mathbf{y}, \theta|X, b) = \exp\left(-\frac{1}{2}||\mathbf{y} - X\theta||_2^2 - \frac{1}{2b}||\theta||_1\right),$$

Spike and slab prior:

$$p(\theta) = \lambda\mathcal{N}(0, \sigma_0) + (1-\lambda)\mathcal{N}(0, \sigma_1),$$

Horseshoe... prior $p(\theta) = \mathcal{N}(0, \lambda)$, $p(\lambda) = Cauchy(0, \tau)$, $p(\tau) = Cauchy(0, 1)$

Comparison on toy example



Large scale data

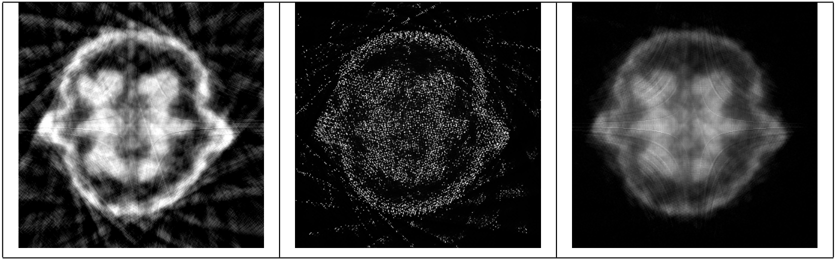
Computer Tomography

$$\mathbf{y} = X\theta + e.$$

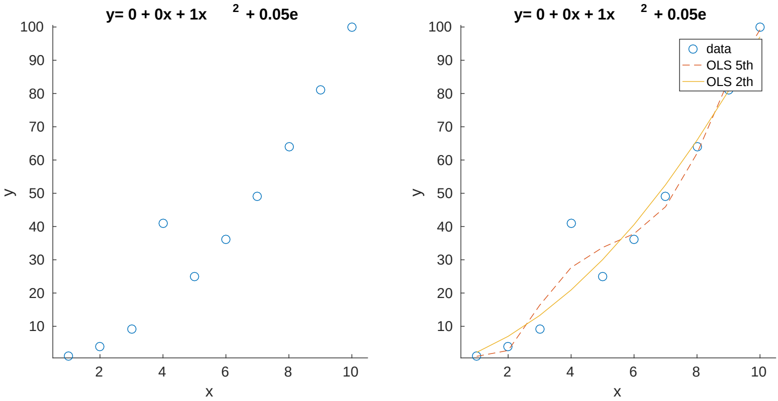
Variational Bayes with full covariance matrix no longer possible.
Maximum of marginal likelihood:

$$\theta^* = \arg\min_\theta \left(\frac{1}{2}\beta||\mathbf{y} - X\theta||_2^2 + \sum_{i=1}^{N} \frac{\nu_i + 1}{2} \ln\left(1 + \frac{\theta_i^2}{\nu_i\sigma_i^2}\right)\right).$$

Non-convex optimization. Prior does matter (at very low numbers of projections).

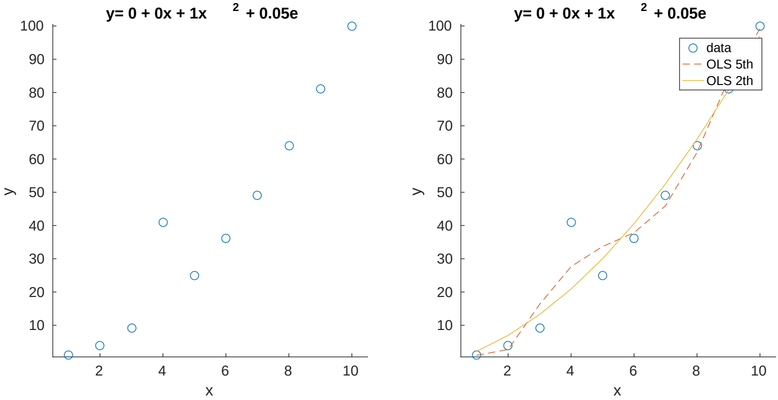


Outliers



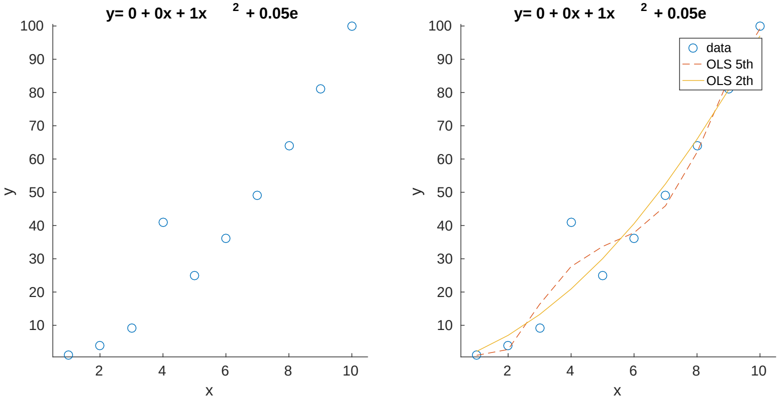
- how to minimize the effect of an outlier?

Outliers



- how to minimize the effect of an outlier?
- outlier detection, robust statistics, etc.

Outliers



- how to minimize the effect of an outlier?
- outlier detection, robust statistics, etc.
- Hierarchical model?

Outliers hierarchical model

Probability model

$$\begin{aligned} p(\mathbf{y}, \theta | X, \alpha) &= p(\mathbf{y} | \theta, X) p(\theta | \alpha) \\ &= \mathcal{N}(X\theta, \beta^{-1} I) \mathcal{N}(0, \alpha^{-1} I). \end{aligned}$$

Outliers hierarchical model

Probability model

$$\begin{aligned} p(\mathbf{y}, \theta | X, \alpha) &= p(\mathbf{y} | \theta, X) p(\theta | \alpha) \\ &= \mathcal{N}(X\theta, \beta^{-1} I) \mathcal{N}(0, \alpha^{-1} I). \end{aligned}$$

Is the variance of the noise homogenous?

Outliers hierarchical model

Probability model

$$
\begin{aligned}
p(\mathbf{y}, \theta | X, \alpha) &= p(\mathbf{y}|\theta, X)p(\theta|\alpha) \\
&= \mathcal{N}(X\theta, \beta^{-1}I)\mathcal{N}(0, \alpha^{-1}I).
\end{aligned}
$$

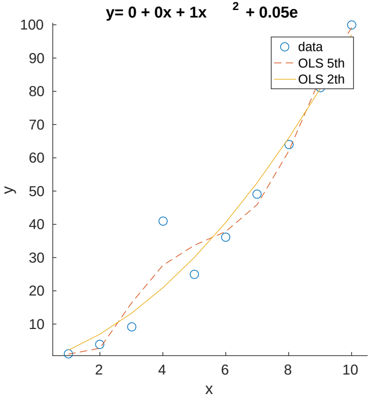
Is the variance of the noise homogenous?
New model:

$$
\begin{aligned}
p(\mathbf{y}, \theta | X, \alpha) &= p(\mathbf{y}|\theta, X)p(\theta|\alpha) \\
&= \mathcal{N}(X\theta, \mathrm{diag}[\beta_1, \dots, \beta_n])\mathcal{N}(0, \alpha^{-1}I).
\end{aligned}
$$

Prior

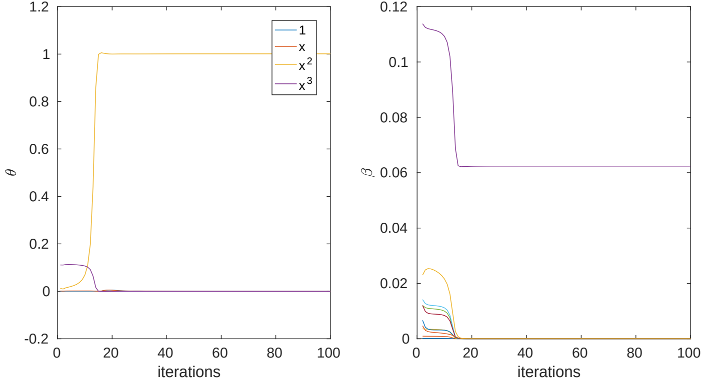
$$
p(\beta_i) = G(\delta, \gamma), \;\; p(\boldsymbol{\beta}) = \prod_i p(\beta_i)
$$

Outliers



y= 0 + 0x + 1x 2 + 0.05e
data
OLS 5th
OLS 2th
100
90
80
70
60
50
40
30
20
10
y
2
4
6
8
10
x
y= 0 + 0x + 1x 2 + 0.05e
data
VB-OLS
120
100
80
60
40
20
0
y
0
2
4
6
8
10
x

Outliers, both diagonal α and β

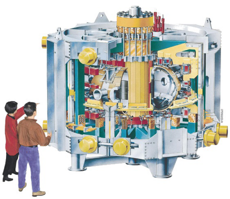


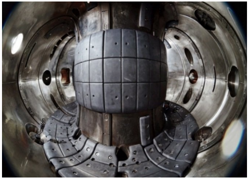
- local minima, unstable for both α and β.

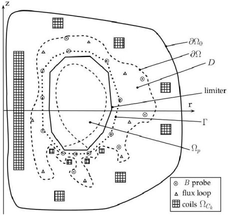
Conclusion

- Linear regression is solved by OLS.
- When the data are not informative, we need to regularize:
- Different prior assumptions yield different results
  - ridge regression minimizes coefficients
  - sparsity prior minimizes the number of non-zero coefficients
- Non-Gaussian residues
  - Student-t residue,
  - Mixture residue, etc.

Assignment: Tokamak plasma boundary







Measurements on loops&coils can be computed by a sum of M_* terms:

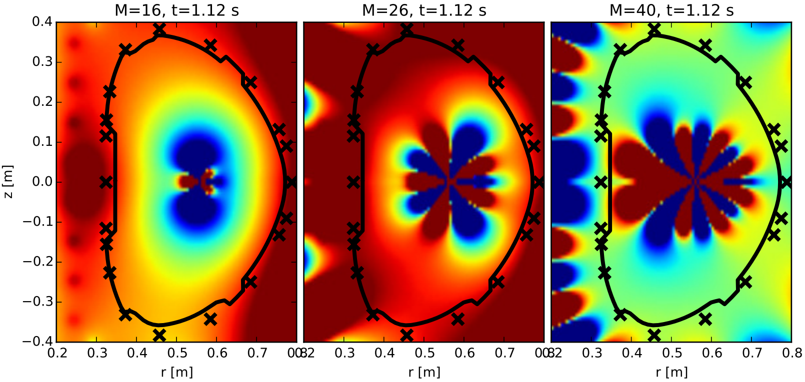
$$\hat{\psi}_{ext}(\zeta,\eta) = \frac{r_0 \sinh\zeta}{\sqrt{\cosh\zeta - \cos\eta}} \times \left\{ \sum_{n=0}^{M_{ea}} a_n^e Q_{n-1/2}^1(\cosh\zeta)\cos(n\eta) + \sum_{n=1}^{M_{eb}} b_n^e Q_{n-1/2}^1(\cosh\zeta)\sin(n\eta) \right\},$$

$$\hat{\psi}_{int}(\zeta,\eta) = \frac{r_0 \sinh\zeta}{\sqrt{\cosh\zeta - \cos\eta}} \times \left\{ \sum_{n=0}^{M_{ia}} a_n^i P_{n-1/2}^1(\cosh\zeta)\cos(n\eta) + \sum_{n=1}^{M_{ib}} b_n^i P_{n-1/2}^1(\cosh\zeta)\sin(n\eta) \right\}.$$

Forming a linear problem $y = X\theta$.

What coefficients are significant?

Noisy observations. OLS not reliable:



File: xy.mat with order 20.

$$y = X\theta + e$$

Find $\hat{\theta}$.

Points

	points
Ridge regression with choice of α	5
Coefficient selection using prior (e.g. ARD)	10
Outlier detection (e.g. ARD)	10