Models with sparse parameters

Václav Šmídl

March 21, 2022

Recapitulation

- maximum likelihood is computed transformation of the likelihood.
  - What transformation we minimize?
  - How it acts on i.i.d. observations?
- Is it possible to interpret an arbitrary loss function as probability?
- What are the interpretations of the additional αI term in the ridge regression $(X^T X + \alpha I)^{-1}$?

Sparse regression: definition

Find a model of data using linear regression, without known regressors

$$y = \begin{cases} a + bx + cx^2 + dx^3 \ldots & \text{polynom} \\ a\exp(x) + b\exp(2x) + c\exp(3x) & \text{exponential} \\ a + bx + c\exp(x) + \ldots & \text{combined} \end{cases}$$

What are the right basis functions?

Sparse regression: definition

Find a model of data using linear regression, without known regressors

$$y = \begin{cases} a + bx + cx^2 + dx^3 \ldots & \text{polynom} \\ a\exp(x) + b\exp(2x) + c\exp(3x) & \text{exponential} \\ a + bx + c\exp(x) + \ldots & \text{combined} \end{cases}$$

What are the right basis functions?
Use as many as possible $X = [1, x, x^2, \exp(x), \exp(2x), \exp(3x), \ldots]$ and solve an optimization problem

$$\hat{\theta} = \arg\min_{\theta}(||y - X\theta||_2^2)$$

$$\text{subject to:} ||\theta||_0 = \sum \theta^0 = m$$

where $m$is a predefined number.

Optimization-based approach

Solving an optimization problem

$$\hat{\theta} = \arg\min_{\theta}(||y - X\theta||_2^2)$$

$$\text{subject to:} ||\theta||_0 = \sum_i^d \theta^0 = m \leq d$$

using the Lagrange multipliers

Optimization-based approach

Solving an optimization problem

$$\hat{\theta} = \arg\min_{\theta}(||y - X\theta||_2^2)$$

$$\text{subject to:} ||\theta||_0 = \sum_i^d \theta^0 = m \leq d$$

using the Lagrange multipliers

$$\hat{\theta} = \arg\min_{\theta} \left(||y - X\theta||_2^2 + \lambda(||\theta||_0 - m) \right).$$

- How to find λ?

Optimization-based approach

Solving an optimization problem

$$\hat{\theta} = \arg\min_{\theta}(||y - X\theta||_2^2)$$

$$\text{subject to:} ||\theta||_0 = \sum_i^d \theta^0 = m \leq d$$

using the Lagrange multipliers

$$\hat{\theta} = \arg\min_{\theta} \left(||y - X\theta||_2^2 + \lambda(||\theta||_0 - m)\right).$$

- How to find λ? Grid search: $\hat{\theta} = \arg\min_{\theta} \left(||y - X\theta||_2^2 + \lambda||\theta||_0\right)$

Optimization-based approach

Solving an optimization problem

$$\hat{\theta} = \arg\min_{\theta}(||y - X\theta||_2^2)$$

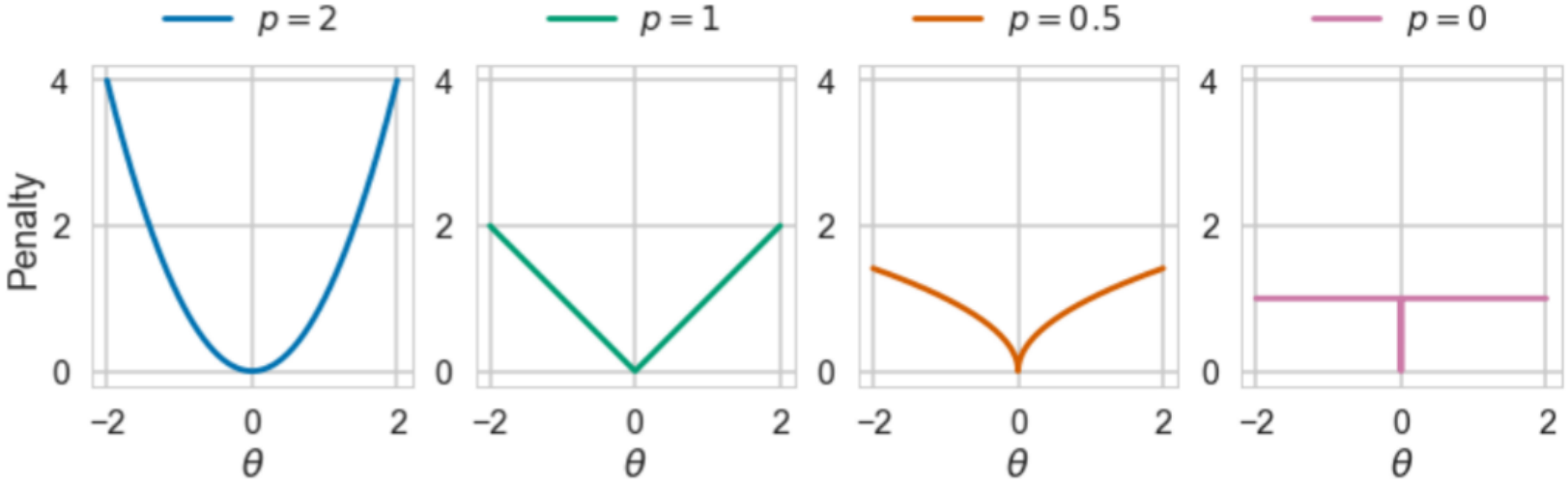
$$\text{subject to:} ||\theta||_0 = \sum_i^d \theta^0 = m \leq d$$

using the Lagrange multipliers

$$\hat{\theta} = \arg\min_{\theta} \left(||y - X\theta||_2^2 + \lambda(||\theta||_0 - m)\right).$$

- How to find λ? Grid search: $\hat{\theta} = \arg\min_{\theta} \left(||y - X\theta||_2^2 + \lambda||\theta||_0\right)$

Hard optimization problem: Relaxation using general L_p norm, $p > 0$



Statistical interpretation via the maximum likelihood approach

Maximum likelihood/aposteriori approach:

$$\hat{\theta}_{ML} = \arg\min_{\theta} \mathcal{L}(\theta) \qquad \mathcal{L}(\theta) = -\log p(\boldsymbol{y}|X, \theta)$$

$$\hat{\theta}_{MAP} = \arg\min_{\theta} \mathcal{L}(\theta) + \mathcal{R}(\theta) \qquad \mathcal{R}(\theta) = -\log p(\theta)$$

Joint Gaussian likelihood with L_1 norm corresponds to the Laplace prior (LASSO)

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

Tight connection between optimization-based and Bayesian.

- optimization corresponds to ML/MAP
- Bayesian can be obtained from optimization via energy model

$$p(\theta) \propto \exp(-\mathcal{L}(\theta))$$

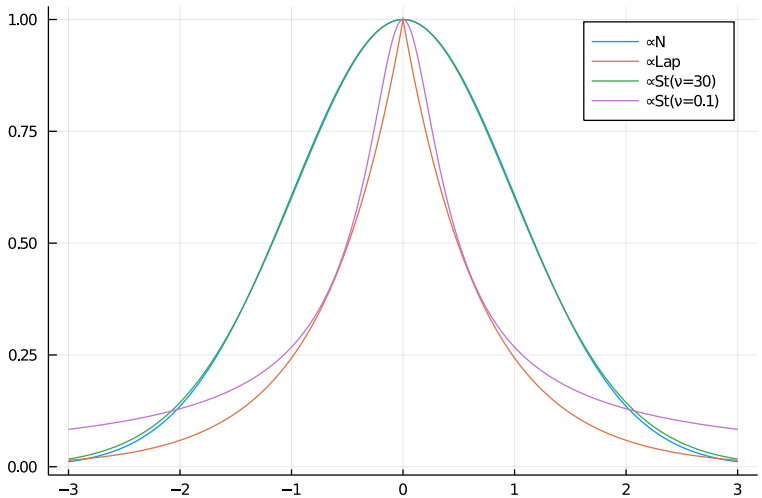
where the challenge is to find the normalization constant.

Beyond Lp norms: what we really want?

The whole purpose is to favor sparse solutions -> increase the prior probability of zeros in the solution.
Lp norm corresponds to a generalized Gaussian distribution
$\exp\left(-\frac{1}{2}||(y-\mu)/\sigma||^p\right)$

Spike and slab prior:

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



Laplace prior

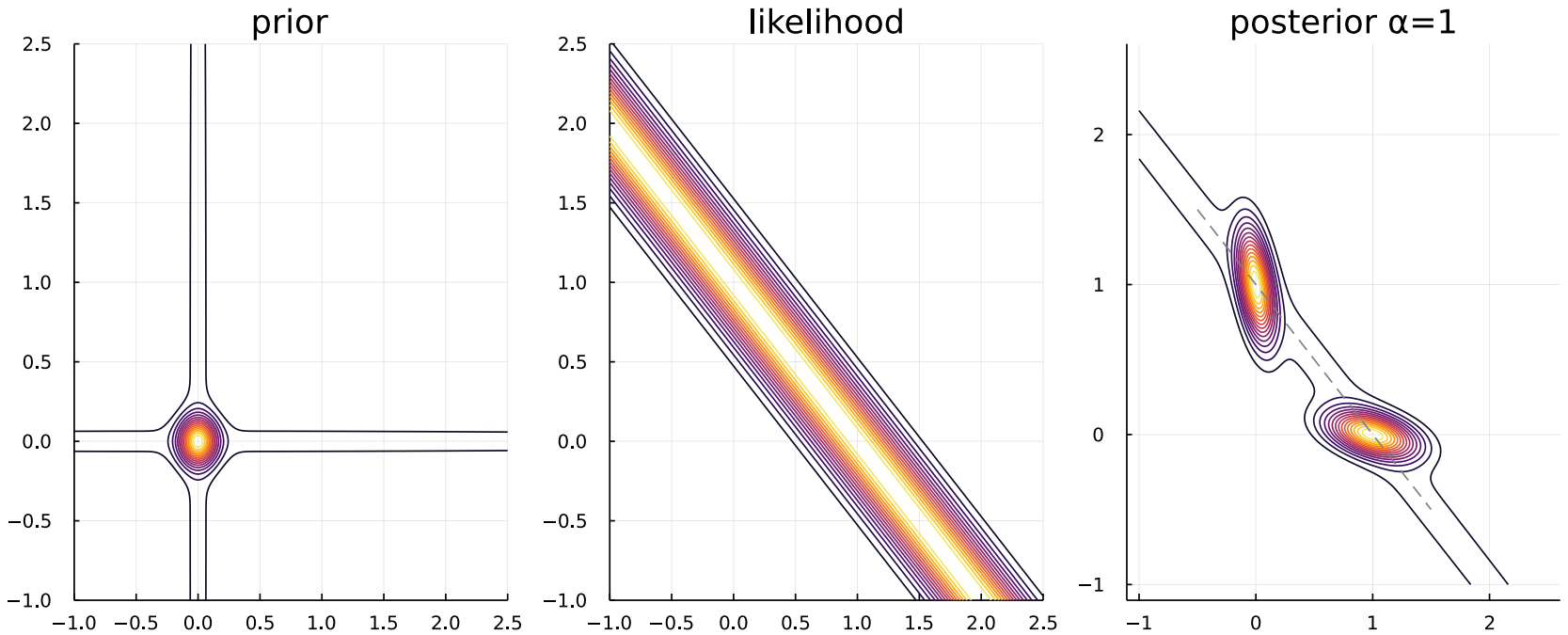
$$p(\theta) = (2b)^{-1}\exp\left(-\frac{1}{2b}\,|x|\right).$$

Student't prior ($\nu \ll 20$):

$$p(\theta) = St(0,\sigma,\nu) = \left(1+\frac{\theta^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}.$$

Spike & slab prior

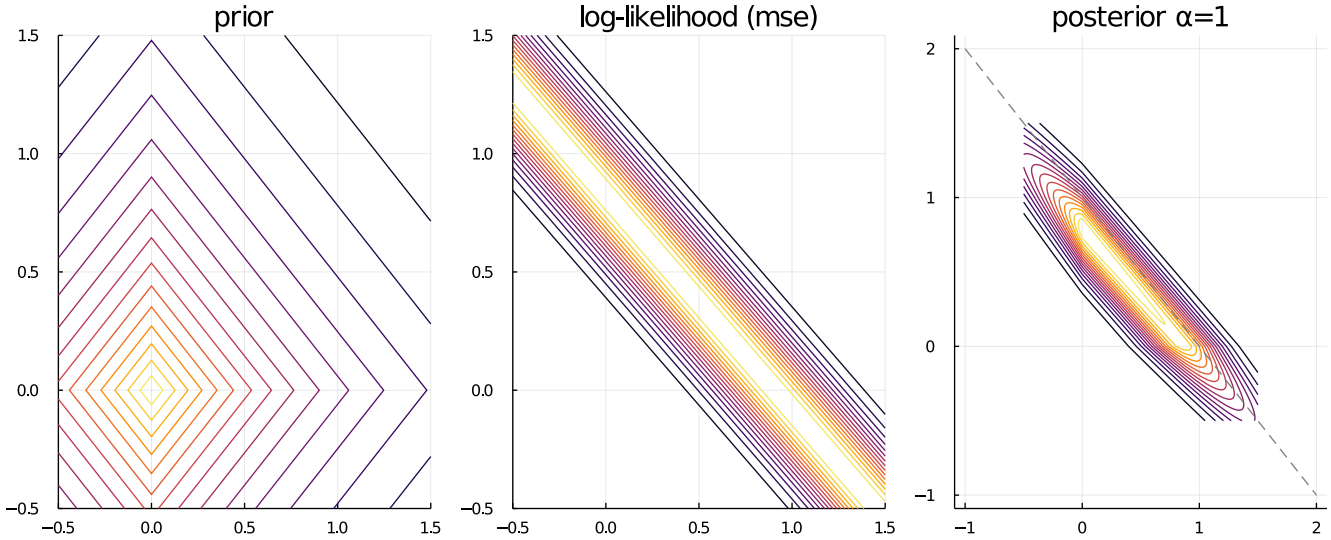
The case of two Gaussians. Technically, the spike can be a Dirac density.



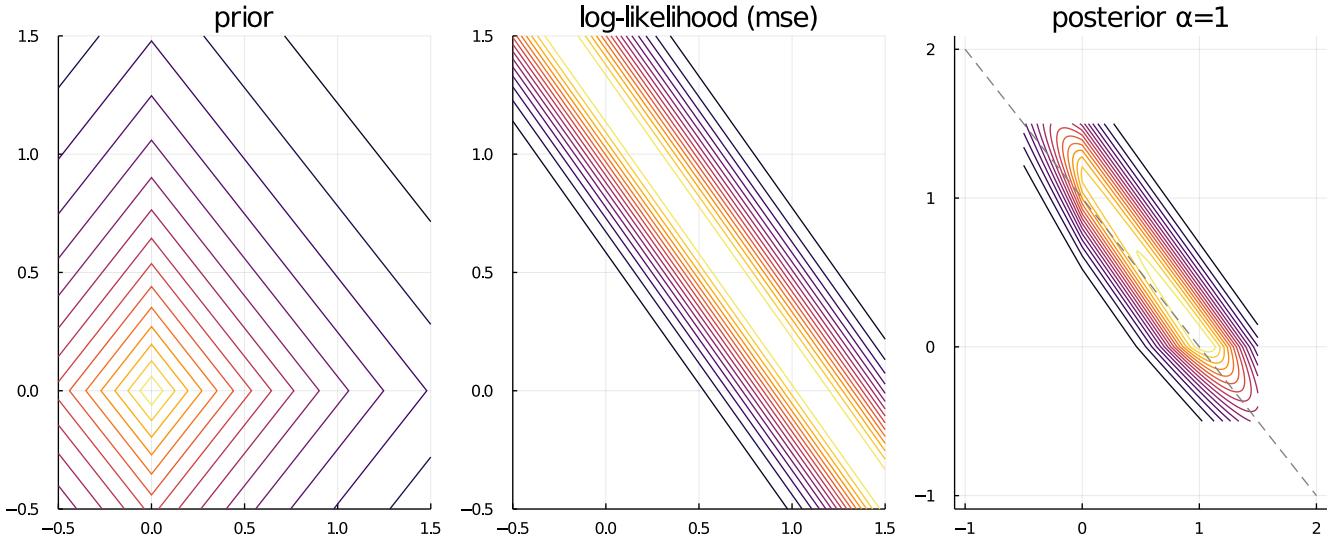
- solution is not unique (2^d extremes)

Uniqueness of Lasso: scaling issue

$x = [0, 1.1], y = [0, 1] \quad \Rightarrow \quad y = x$



$x = [0, 0.9], y = [0, 1] \quad \Rightarrow \quad y = x^2$



Gaussian scale mixture

Is a general class of models of the form

$$p(\theta|\sigma_\theta) = \mathcal{N}(0, \sigma_\theta), \qquad p(\sigma_\theta) \qquad p(\theta) = \int p(\theta|\sigma_\theta)p(\sigma_\theta)d\sigma_\theta$$

with special cases

$$p(\sigma_\theta) = w_1\delta(\sigma_\theta - \sigma_1) + w_2\delta(\sigma_\theta - \sigma_2), \quad p(\theta) = w_1\mathcal{N}(0, \sigma_1) + w_2\mathcal{N}(0, \sigma_2)$$

$$p(\sigma_\theta) = i\Gamma(\alpha_0, \beta_0), \qquad p(\theta) = \mathcal{S}t\left(2\alpha_0, 0, (\alpha\beta)^{-1/2}\right)$$

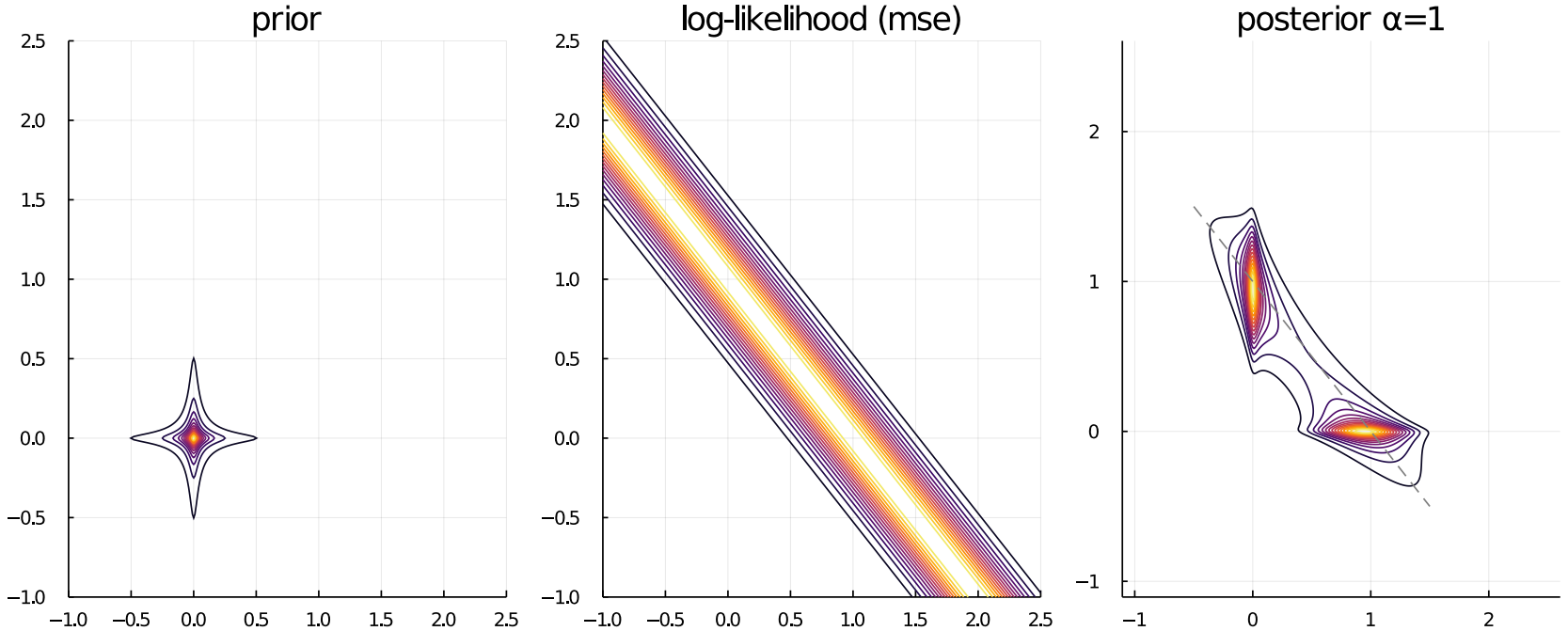
Also available in multiplicative version:

$$\theta = wz, \quad p(w|\sigma_0) = \mathcal{N}(0, \sigma_0), \quad p(z) \quad p(\theta) = \int \mathcal{N}(0, \sigma_0 z)p(z)dz$$

with many choises of $p(z)$, e.g. $z \in \{0, 1\}$ an assumption a range

Student't prior, $\mathcal{R} = \log(1+\theta^2/\nu)$

Consider $p(\theta_i) = St(0, \nu, \sigma_0), i = 1, \ldots, p$ i.e. $p(\theta) = \prod_{i=1}^{p} St(0, \nu, \sigma_0)$.
For $\nu = 0.001, \sigma_0 = 1$:



Multiple solutions. Scaling affects which one is global.

- How to do Bayesian inference?
- Student't is not conjugate.

The power of latent variable

Linear regression with Gaussian scale mixture prior:

$$\begin{aligned} p(\boldsymbol{y}|X,\theta,\omega) &= \mathcal{N}(X\theta,\omega^{-1}I) \propto \omega^{n/2}\exp\left(-\frac{1}{2}\omega(\boldsymbol{y}-X\theta)^T(\boldsymbol{y}-X\theta)\right) \\ p(\theta) &= \prod_i p(\theta_i) = \prod_i p(\theta_i|\sigma_i)p(\sigma_i) \end{aligned}$$

Motivates to replace the Bayes rule:

$$\begin{aligned} p(\theta|X,\boldsymbol{y}) &\propto p(\boldsymbol{y}|X,\theta)\int p(\theta|s)p(s)ds \\ &\propto \int p(\boldsymbol{y}|X,\theta)p(\theta|s)p(s)ds \\ &= \int p(\theta,s|X,\boldsymbol{y})ds \end{aligned}$$

Solving $p(\theta,s|X,\boldsymbol{y})$ e.g. by Variational Bayes $p \approx q(\theta)q(s)$.

Automatic relevance determination (ARD)

Probability model

$$
\begin{aligned}
p(\mathbf{y}, \theta | X, \alpha, \omega) &= \mathcal{N}(X\theta, \omega^{-1} I) \\
p(\theta | \alpha) &= \mathcal{N}(0, \text{diag}[\alpha_1, \ldots, \alpha_p]^{-1}) \\
p(\alpha_i) &= G(\delta_0, \gamma_0), \quad p(\boldsymbol{\alpha}) = \prod_i p(\alpha_i) \\
p(\omega) &= G(0, 0)
\end{aligned}
$$

with joint distribution

$$
p(\mathbf{y}, \theta, \alpha, \omega | X) \propto \exp\left\{ -\frac{1}{2}\omega ||\mathbf{y} - X\theta||_2^2 - \frac{1}{2}\sum_i \alpha_i \theta_i^2 - \sum_i \alpha_i \gamma_0 \right\} \omega^n \prod_i \alpha_i^{\delta_0 +}
$$

$$
\begin{aligned}
\log p(\mathbf{y}, \theta, \alpha, \omega | X) = &-\frac{1}{2}\omega ||\mathbf{y} - X\theta||_2^2 - \frac{1}{2}\sum_i \alpha_i \theta_i^2 - \sum_i \alpha_i \gamma_0 \\
&+ (\delta_0 + 1) \log \alpha_i + \log \omega
\end{aligned}
$$

Variational Bayes for OLS with ARD

Log-joint distribution

$$\log p(\mathbf{y}, \theta, \alpha, \omega | X) = -\frac{1}{2}\omega||\mathbf{y} - X\theta||_2^2 - \frac{1}{2}\sum_i \alpha_i \theta_i^2 - \sum_i \alpha_i \gamma_0 + (\delta_0 + 1)\log \alpha_i + \frac{p}{2}\log \omega$$

Using Variational Bayes (cover-up rule):

$$\log q(\theta | X) = -\frac{1}{2}\omega||\mathbf{y} - X\theta||_2^2 - \frac{1}{2}\sum_i \mathrm{E}(\alpha_i)\,\theta_i^2$$

$$\log q(\alpha_i | X) = -\frac{1}{2}\alpha_i \mathrm{E}(\theta_i^2) - \alpha_i \gamma_0 + (\delta_0 + 1)\log \alpha_i$$

$$\log q(\omega | X) = -\frac{1}{2}\omega \mathrm{E}(||\mathbf{y} - X\theta||_2^2) + \frac{p}{2}\log \omega$$

Variational Bayes for Automatic relevance determination

Probability model

$$p(\mathbf{y}, \theta | X, \alpha, \omega) = \mathcal{N}(X\theta, \omega I)\mathcal{N}(0, \text{diag}[\alpha_1, \ldots, \alpha_p]) \prod_i G(\gamma_0, \delta_0)$$

Posterior factors

$$\begin{aligned}
p(\alpha_i | \mathbf{y}, X) &= G(\gamma, \delta_i), \\
\delta_i &= \delta_0 + \frac{1}{2}\mathsf{E}\left(\theta_i^2\right), \ \gamma = \gamma_0 + \frac{1}{2}, \\
p(\theta | \mathbf{y}, X) &= \mathcal{N}(\hat{\theta}, \Sigma_\theta), \\
\hat{\theta} &= (\omega X^T X + \text{diag}\mathsf{E}(\alpha))^{-1} \omega X^T \mathbf{y}, \\
\Sigma_\theta &= (\omega X^T X + \text{diag}\mathsf{E}(\alpha))^{-1}. \\
p(\omega | \boldsymbol{y}, X) &= G(p/2, \mathsf{E}((\boldsymbol{y} - X\theta)^T (\boldsymbol{y} - X\theta)))
\end{aligned}$$

Iterated least squares. Moment

Variational Bayes for Automatic relevance determination

Probability model

$$p(\mathbf{y}, \theta | X, \alpha, \omega) = \mathcal{N}(X\theta, \omega I)\mathcal{N}(0, \text{diag}[\alpha_1, \ldots, \alpha_p]) \prod_i G(\gamma_0, \delta_0)$$

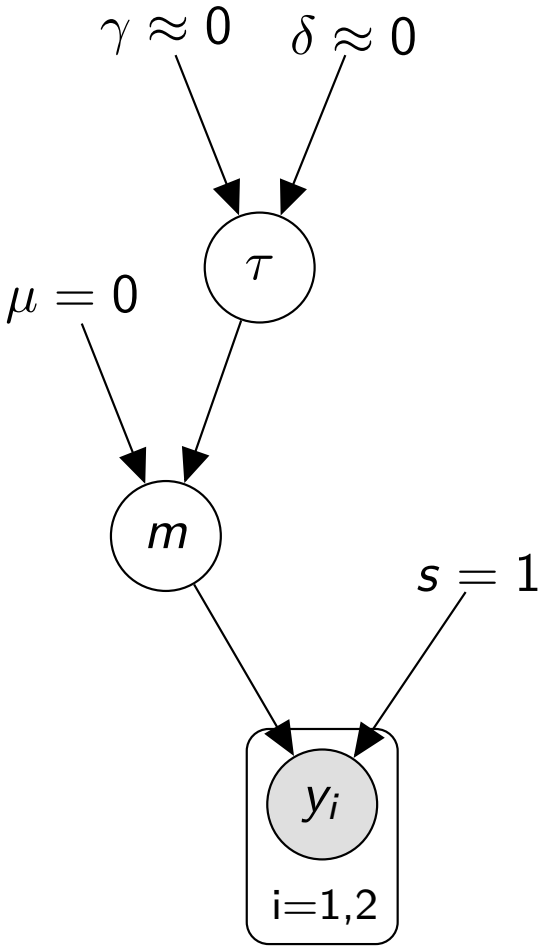
Posterior factors

$$\begin{aligned}
p(\alpha_i | \mathbf{y}, X) &= G(\gamma, \delta_i), \\
\delta_i &= \delta_0 + \frac{1}{2}\mathsf{E}\left(\theta_i^2\right), \ \gamma = \gamma_0 + \frac{1}{2}, \\
p(\theta | \mathbf{y}, X) &= \mathcal{N}(\hat{\theta}, \Sigma_\theta), \\
\hat{\theta} &= (\omega X^T X + \text{diag}\mathsf{E}(\alpha))^{-1} \omega X^T \mathbf{y}, \\
\Sigma_\theta &= (\omega X^T X + \text{diag}\mathsf{E}(\alpha))^{-1}. \\
p(\omega | \boldsymbol{y}, X) &= G(p/2, \mathsf{E}((\boldsymbol{y} - X\theta)^T(\boldsymbol{y} - X\theta)))
\end{aligned}$$

Iterated least squares. Moment

$$\mathsf{E}\left(\theta_i^2\right) = \hat{\theta}_i^2 + \Sigma_{\theta,i,i}, \quad \mathsf{E}(\alpha_i) = \frac{\gamma}{\delta_i}, \quad \mathsf{E}(\theta^T M \theta) = \hat{\theta}^T M \hat{\theta} + \text{tr}(M\Sigma_\theta)$$

Revisiting the toy example:

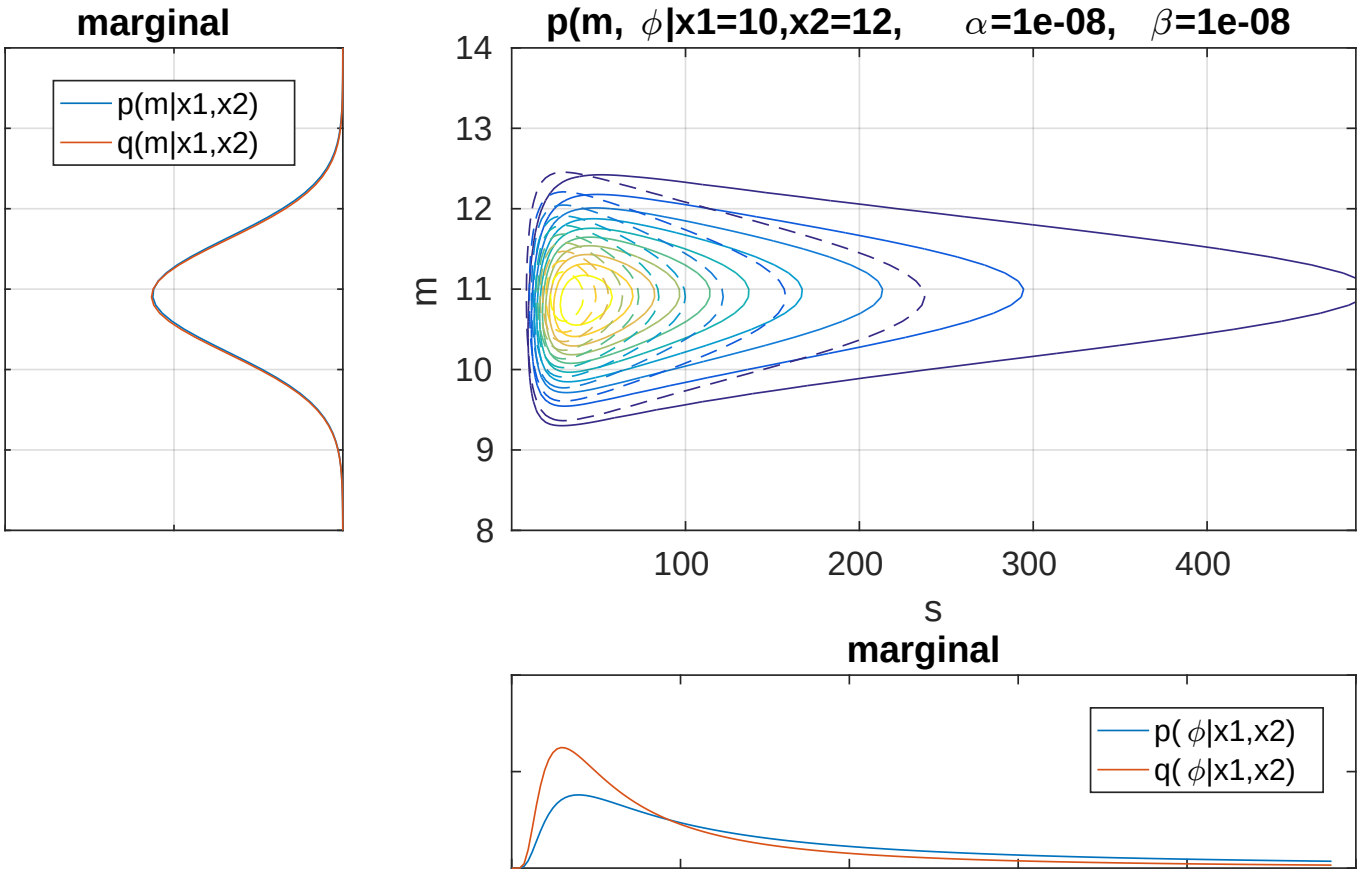


$$
\begin{aligned}
p(y_i|m) &= \mathcal{N}(m, 1), \\
p(m|\tau) &= \mathcal{N}(0, \tau), \\
p(\tau) &= iG(\gamma_0, \delta_0)
\end{aligned}
$$

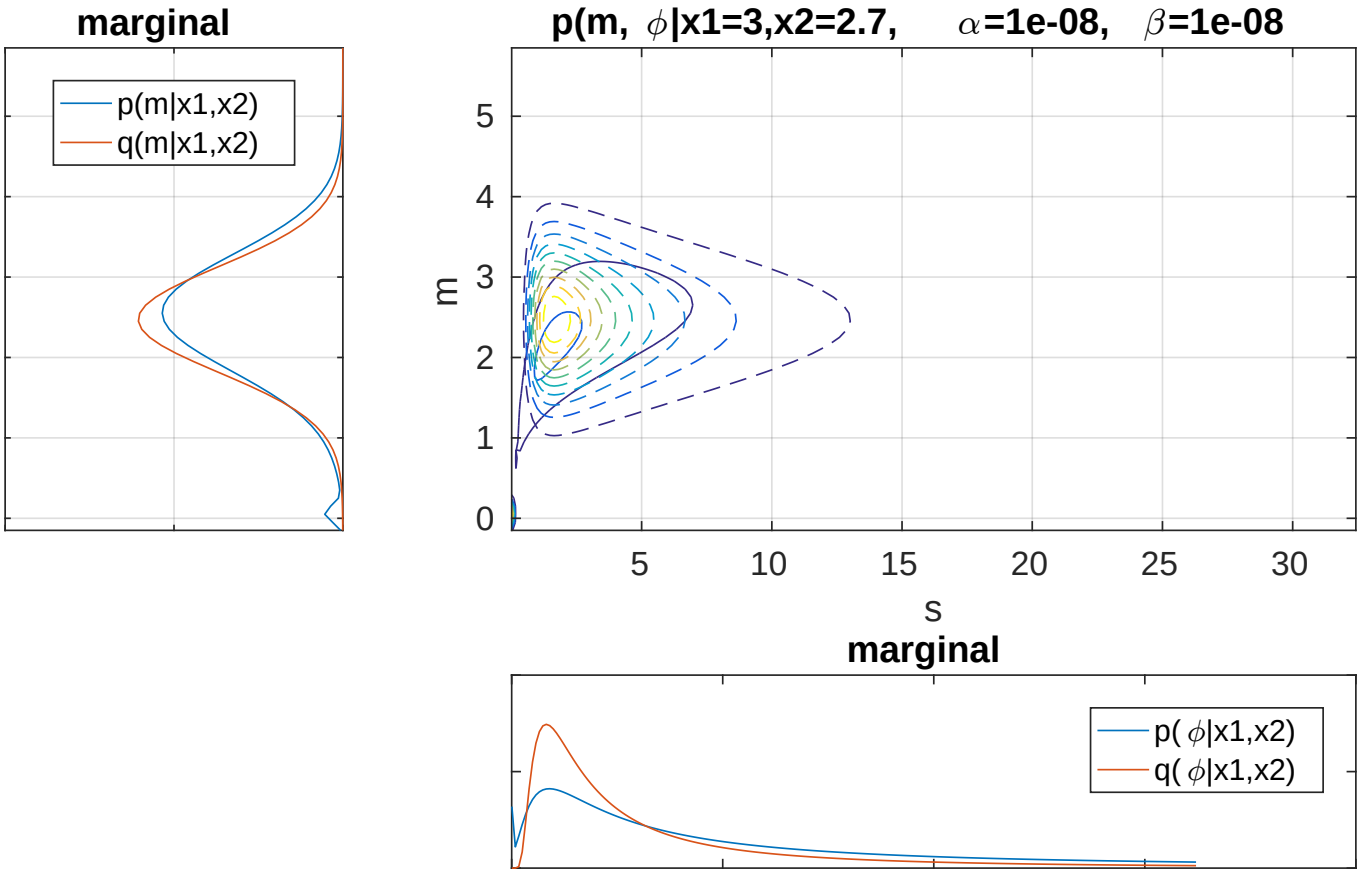
Solution using:

- Variational Bayes
- Numerical evaluation on grid

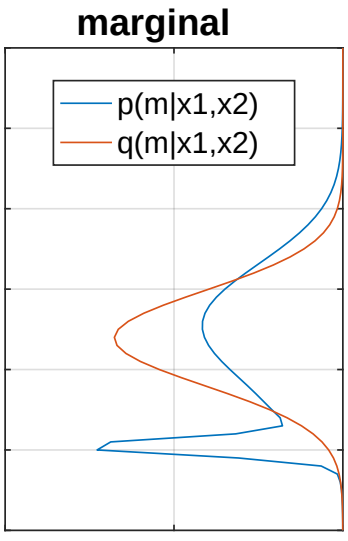
Revisiting toy example:

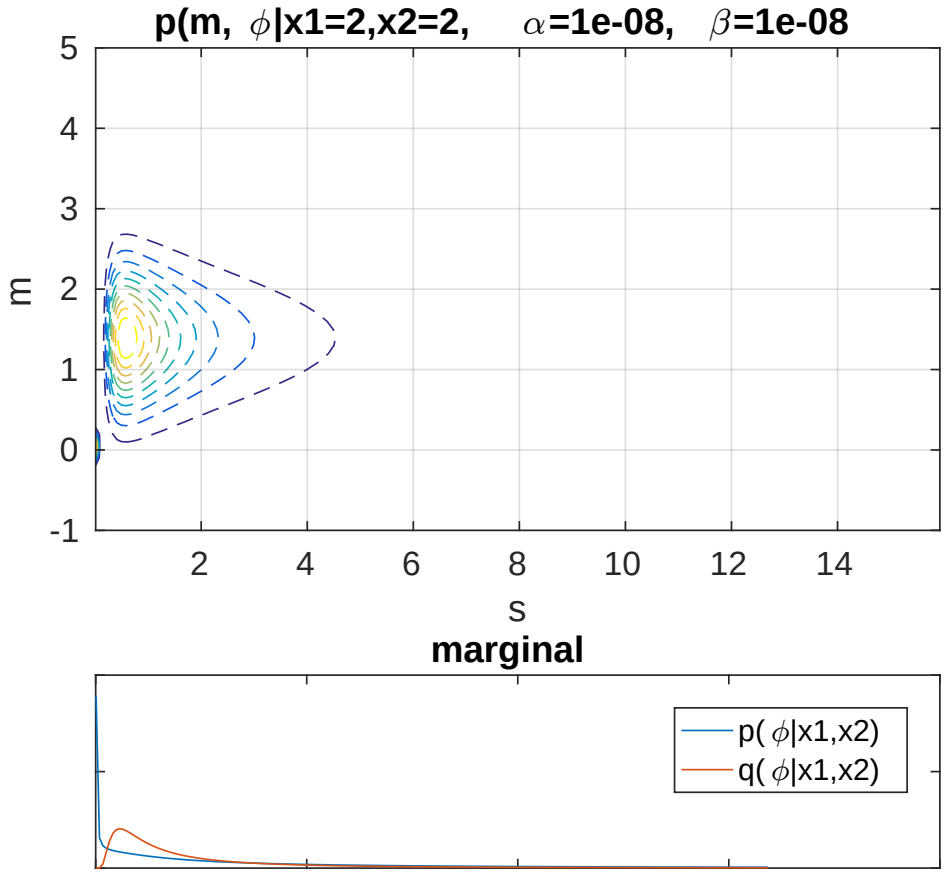


Revisiting toy example:

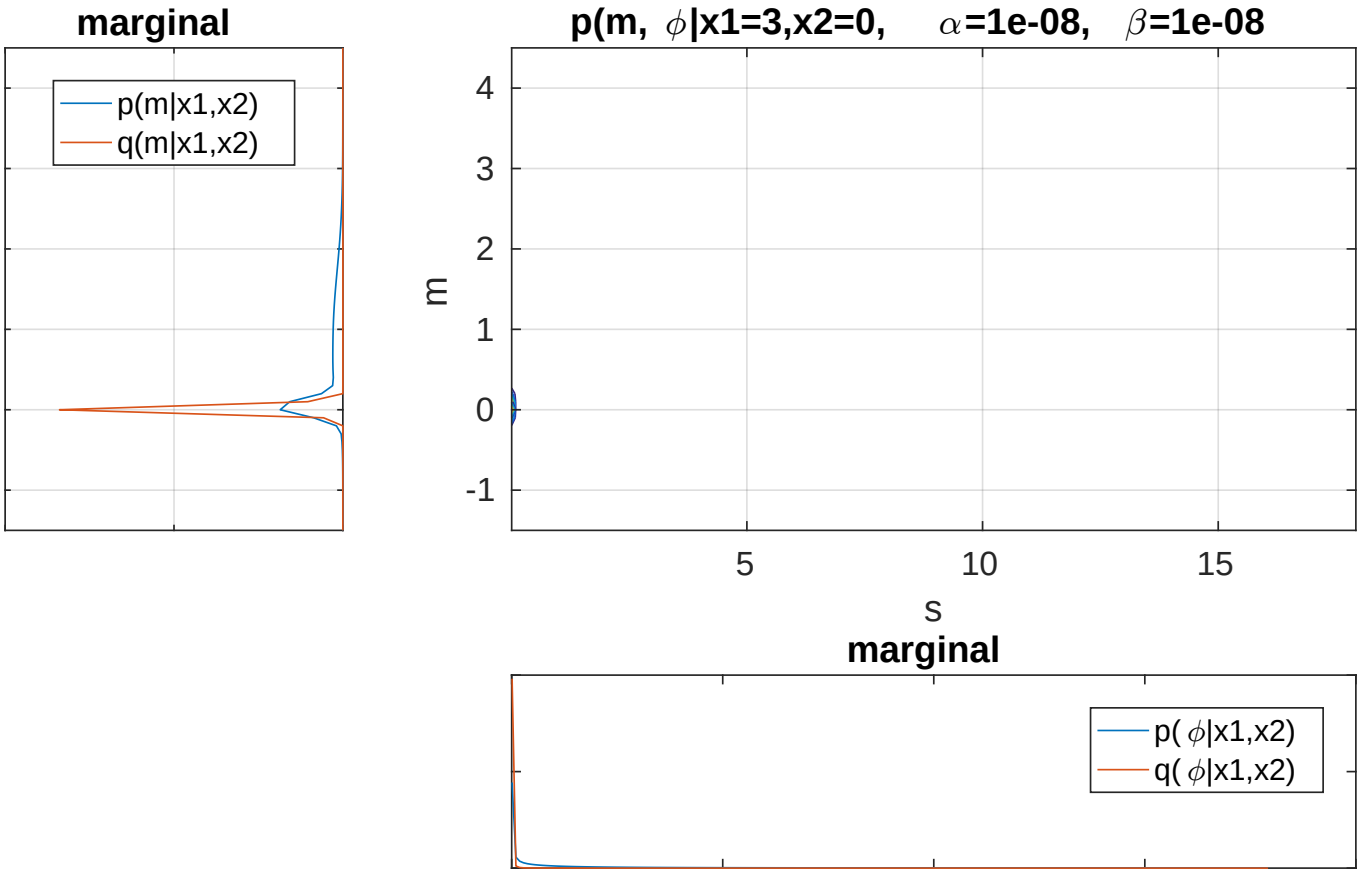


Revisiting toy example:

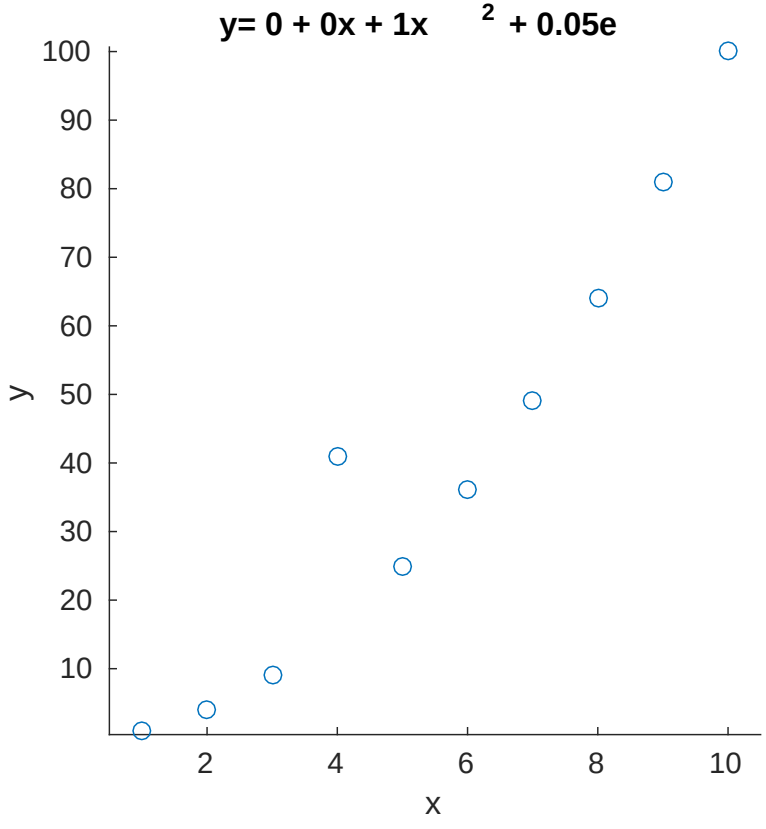


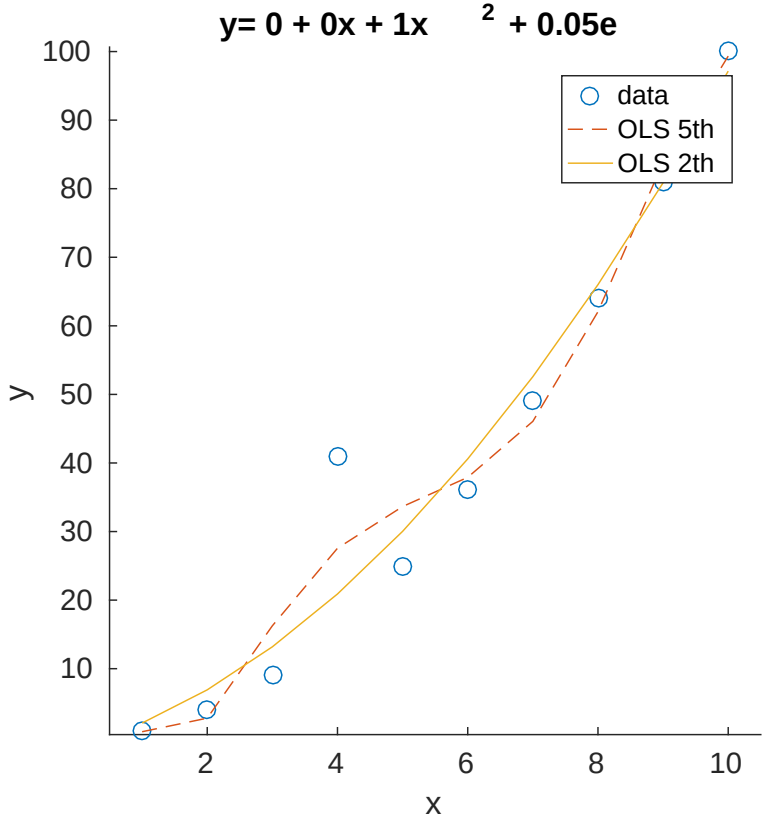


Revisiting toy example:



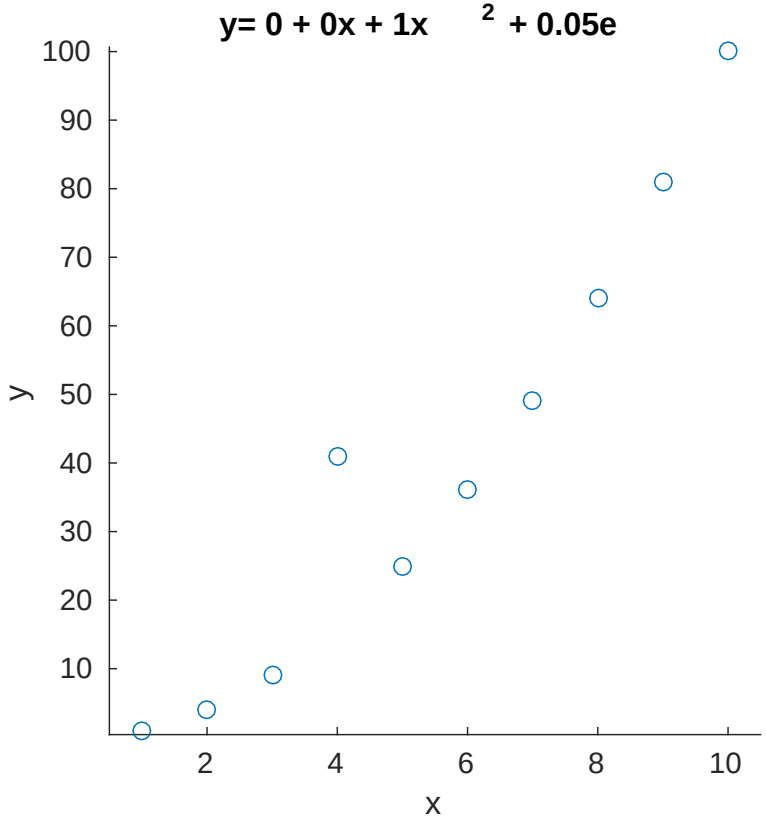
Outliers

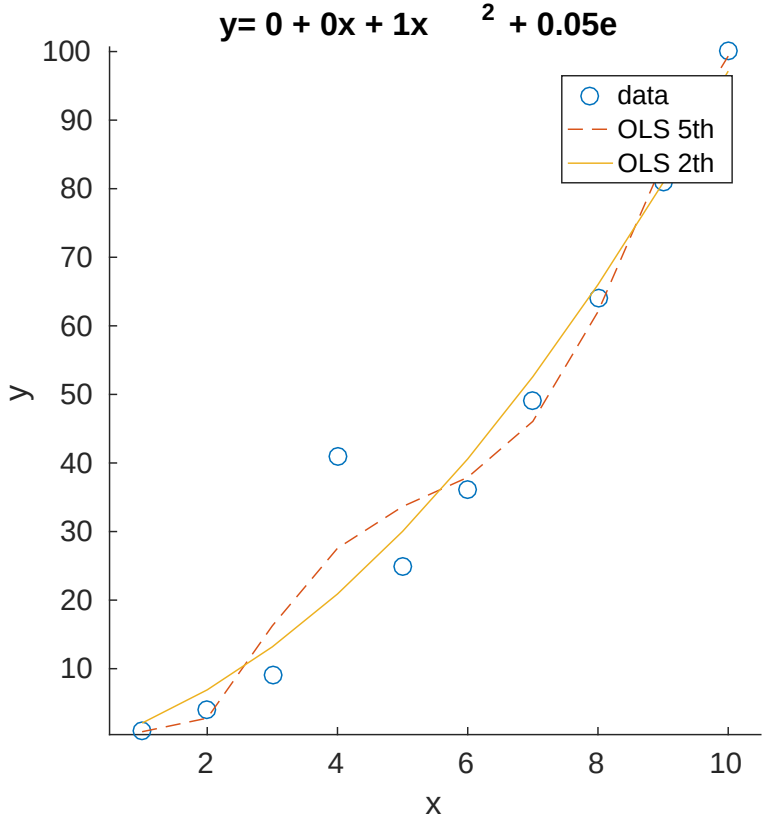




- How to minimize the effect of an outlier?

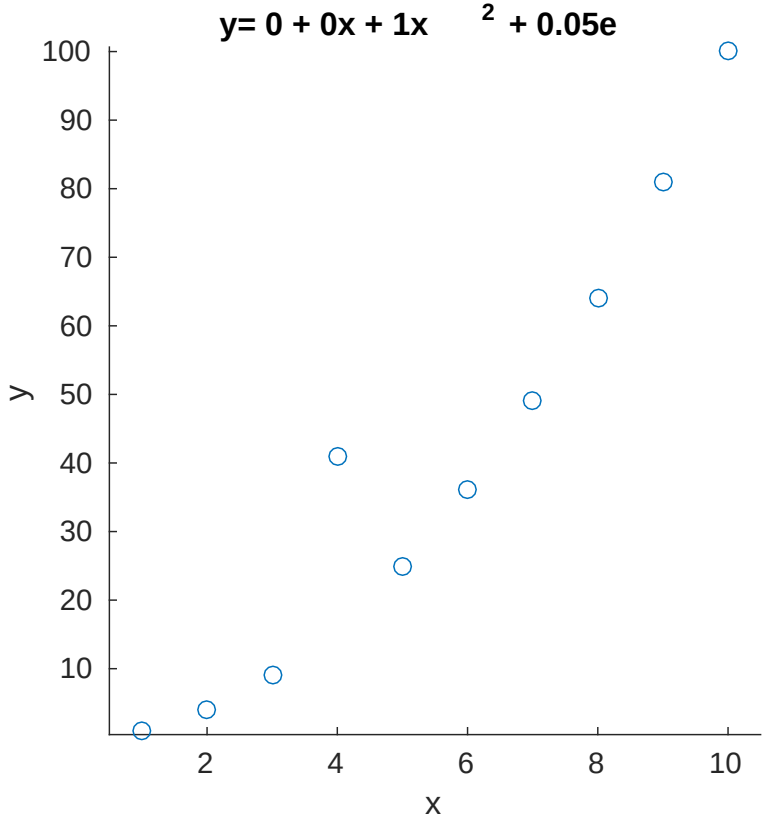
Outliers

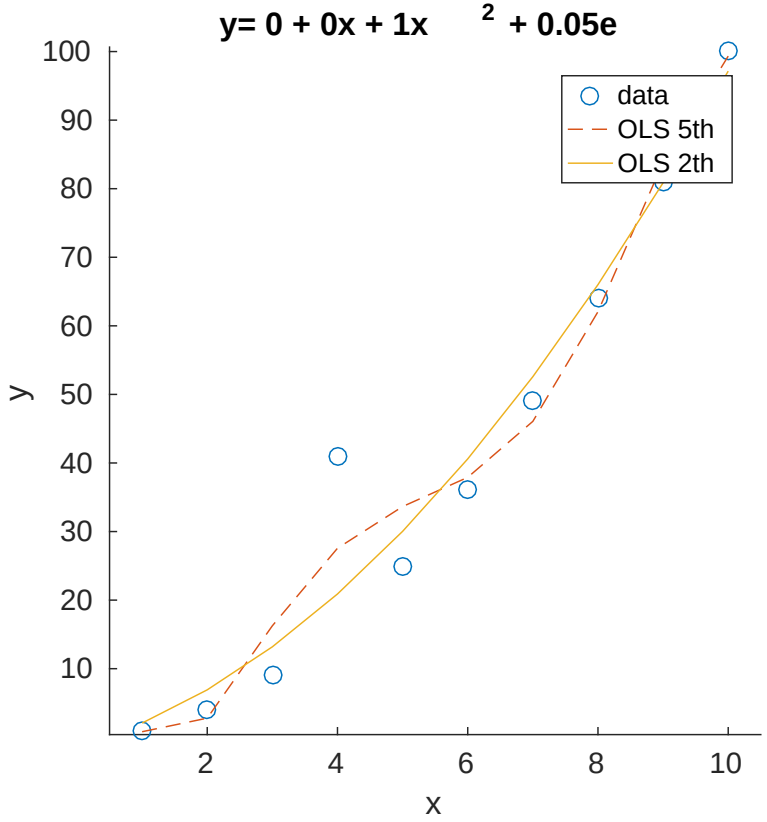




- How to minimize the effect of an outlier?
- Outlier detection, robust statistics, etc.

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

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

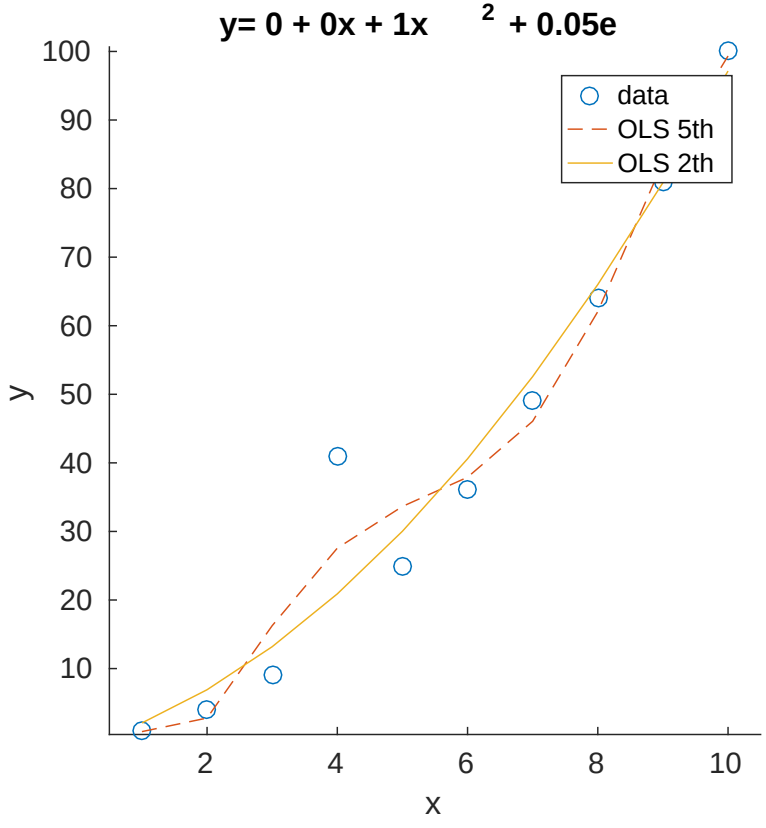
Is the variance of the noise homogenous?
New model:

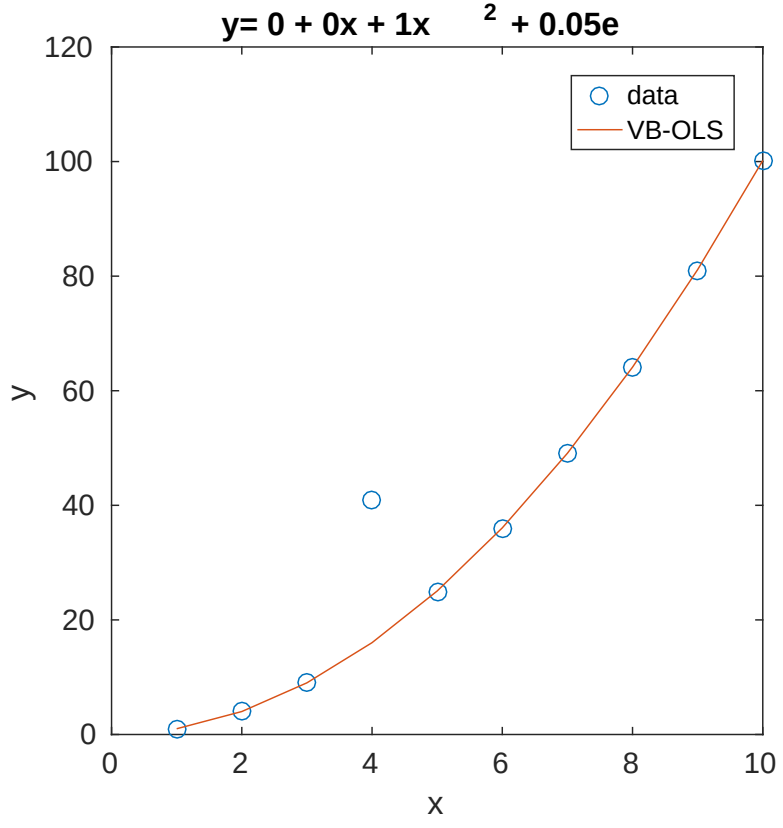
$$\begin{aligned} p(\mathbf{y}, \theta | X, \alpha) &= p(\mathbf{y}|\theta, X) p(\theta|\alpha) \\ &= \mathcal{N}(X\theta, \text{diag}[\beta_1, \ldots, \beta_n]) \mathcal{N}(0, \alpha^{-1} I). \end{aligned}$$

Prior

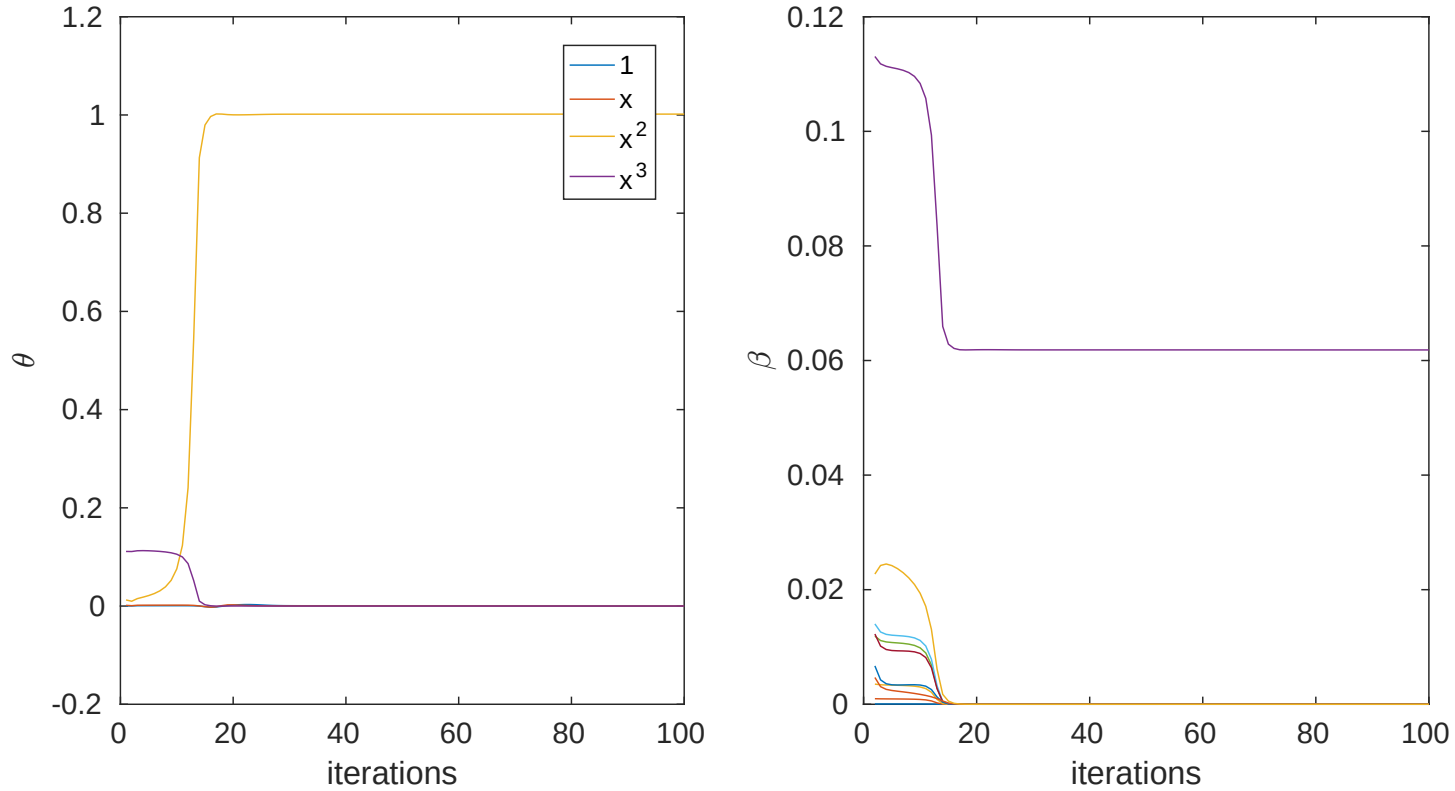
$$p(\beta_i) = G(\delta, \gamma), \;\; p(\boldsymbol{\beta}) = \prod_i p(\beta_i)$$

Outliers


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


y= 0 + 0x + 1x 2 + 0.05e
data
VB-OLS
y
x
120
100
80
60
40
20
0
0
2
4
6
8
10

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.

ETEX data challenge

Estimate source term of the ETEX experiment

$$y = X\theta,$$

where θ is assumed to be sparse, piece-wise linear, with non-negative elements.

	points
θ using "non-convex" prior (e.g. ARD)	5
Outlier detection (e.g. ARD)	5

Beware:

- scale of the matrix is very low! Normalize to $X/\max(X)$