Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2012
Prof. Erik Sudderth

Lecture 13:
Generalized Linear Models
Robust Regression
Feature Selection & Search

Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

Exponential Families of Distributions

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})]$$
$$= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})]$$

$$Z(\boldsymbol{\theta}) = \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x}$$
$$A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$$

$\phi(x) \in \mathbb{R}^d \longrightarrow$ fixed vector of *sufficient statistics* (features), specifying the family of distributions

$\theta \in \Theta \longrightarrow$ unknown vector of *natural parameters*, determine particular distribution in this family

$Z(\theta) > 0 \longrightarrow$ normalization constant or *partition function*, ensuring this is a valid probability distribution

$h(x) > 0 \longrightarrow$ *reference measure* independent of parameters (for many models, we simply have $h(x) = 1$)

To ensure this construction is valid, we take

$$\Theta = \{\theta \in \mathbb{R}^d \mid Z(\theta) < \infty\}$$

Examples of Exponential Families

$$
\begin{aligned}
p(\mathbf{x}|\boldsymbol{\theta}) &= \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] \qquad & Z(\boldsymbol{\theta}) &= \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x} \\
&= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] \qquad & A(\boldsymbol{\theta}) &= \log Z(\boldsymbol{\theta})
\end{aligned}
$$

- Bernoulli and binomial (2 classes) $\phi(x) = \mathbb{I}(x = 1) = x$
- Categorical and multinomial (K classes)

$$\phi(x) = [\mathbb{I}(x = 1), \ldots, \mathbb{I}(x = K - 1)]$$

- Scalar Gaussian $\phi(x) = [x, x^2]$
- Multivariate Gaussian $\phi(x) = [x, xx^T]$

- Poisson $h(x) = \frac{1}{x!}, \phi(x) = x$

- Dirichlet and beta
- Gamma and exponential
- …

Learning in Exponential Families

$$
\begin{aligned}
p(\mathbf{x}|\boldsymbol{\theta}) &= \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] \\
&= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})]
\end{aligned}
\qquad
\begin{aligned}
Z(\boldsymbol{\theta}) &= \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x} \\
A(\boldsymbol{\theta}) &= \log Z(\boldsymbol{\theta})
\end{aligned}
$$

- For *maximum likelihood* estimation, we find the *unique* set of parameters which satisfy:

$$\mathbb{E}_\theta[\phi(x)] = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i) \qquad \nabla_\theta A(\theta) = \mathbb{E}_\theta[\phi(x)]$$

- Special cases we've seen: *Categorical, Gaussian, …*
- For *Bayesian* estimation, there are convenient properties:
  - Except for a few "odd" exceptions, exponential families are the only distributions with *conjugate priors*
  - Leads to more tractable posteriors and marginal likelihoods
  - There is a simple formula for constructing these priors: *Beta-Bernoulli, Dirichlet-categorical, Gaussian-Gaussian, …*

Generalized Linear Models

- General framework for modeling non-Gaussian data with linear prediction, using exponential families:
  - Construct instance-specific natural parameters:

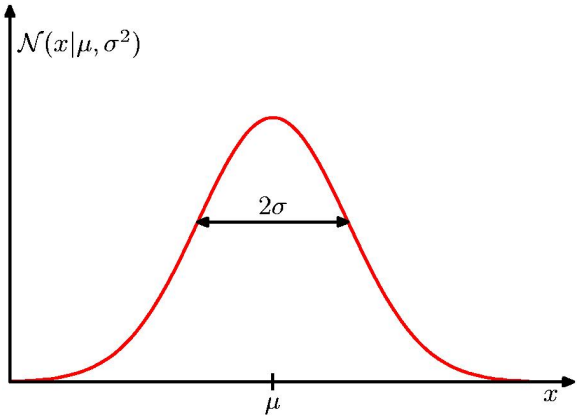
$$\theta_i = w^T \phi(x_i)$$

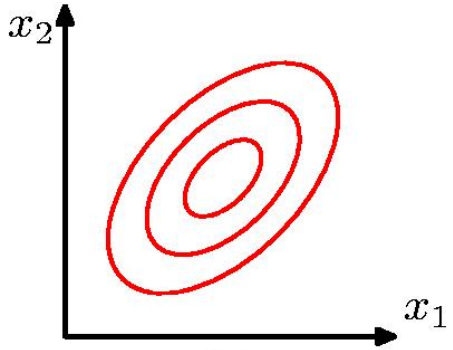
  - Observation comes from exponential family:

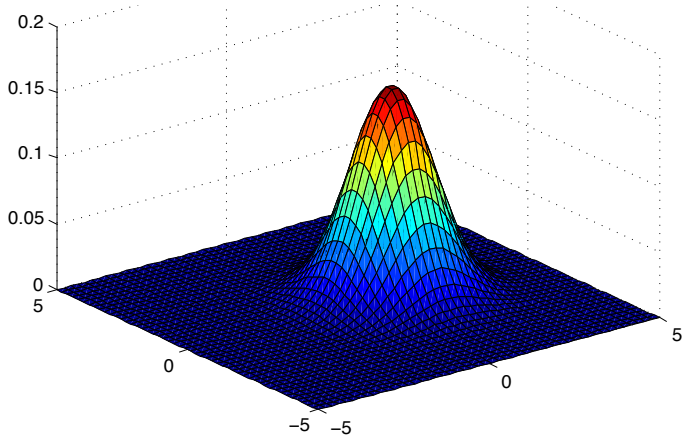
$$p(y_i \mid x_i, w) = \exp\left\{y_i \theta_i - A(\theta_i)\right\}$$

- Special cases: linear regression and logistic regression
- ML and MAP estimation is generally straightforward
- Many possible extensions:
  - *Multivariate responses* with more parameters (biggest difficulty is notation and indexing)
  - *Link functions* to allow more flexibility in how $(w, x_i) \rightarrow \theta_i$

Gaussian Distribution







Why the Gaussian distribution?

- *Central limit theorem:* Property of (some) big datasets
- *Flexibility:* Can capture arbitrary mean and covariance
- *Convenience:* Quadratic log-likelihood easy to optimize

Why consider non-Gaussian likelihoods?

- *Data type:* Observations may not be continuous numbers
- *Outliers:* Increase robustness to non-typical data

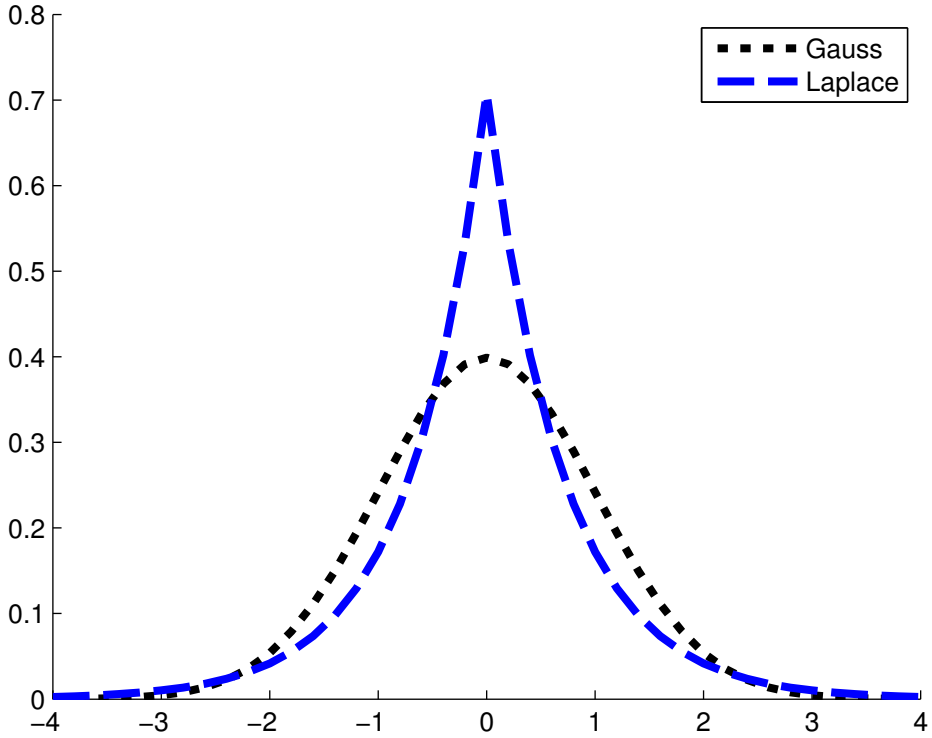
Why consider non-Gaussian priors on parameters?

- *Sparsity:* Allow selection of most important model features

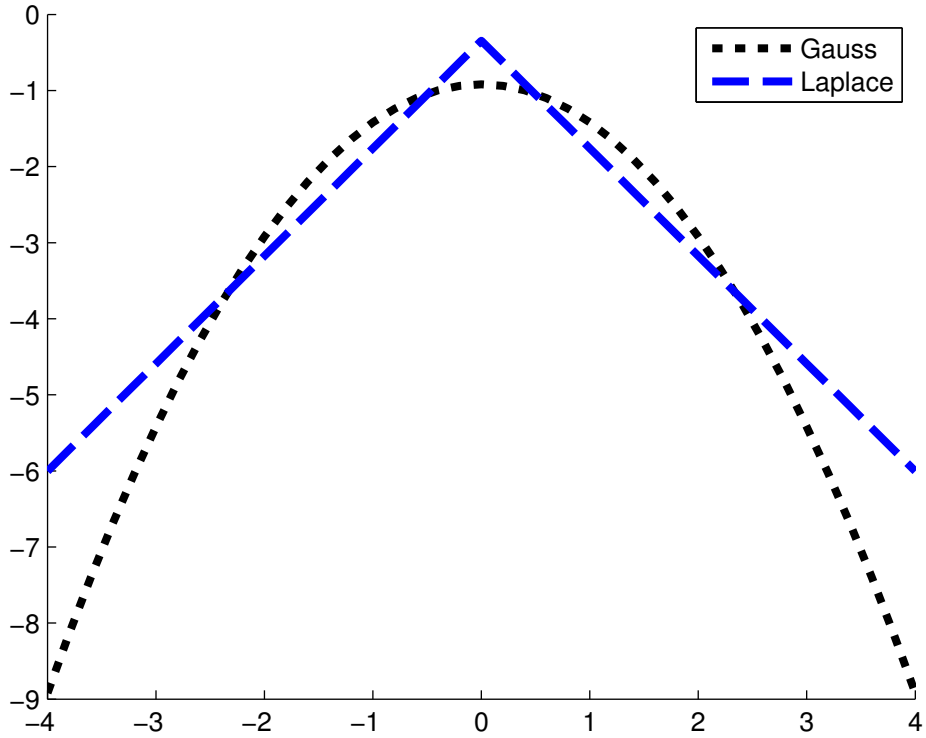
Laplace Distribution

$$\text{Lap}(x \mid \mu, \lambda) = \frac{\lambda}{2} \exp(-\lambda|x - \mu|)$$

Probability Densities



Log Probability Densities



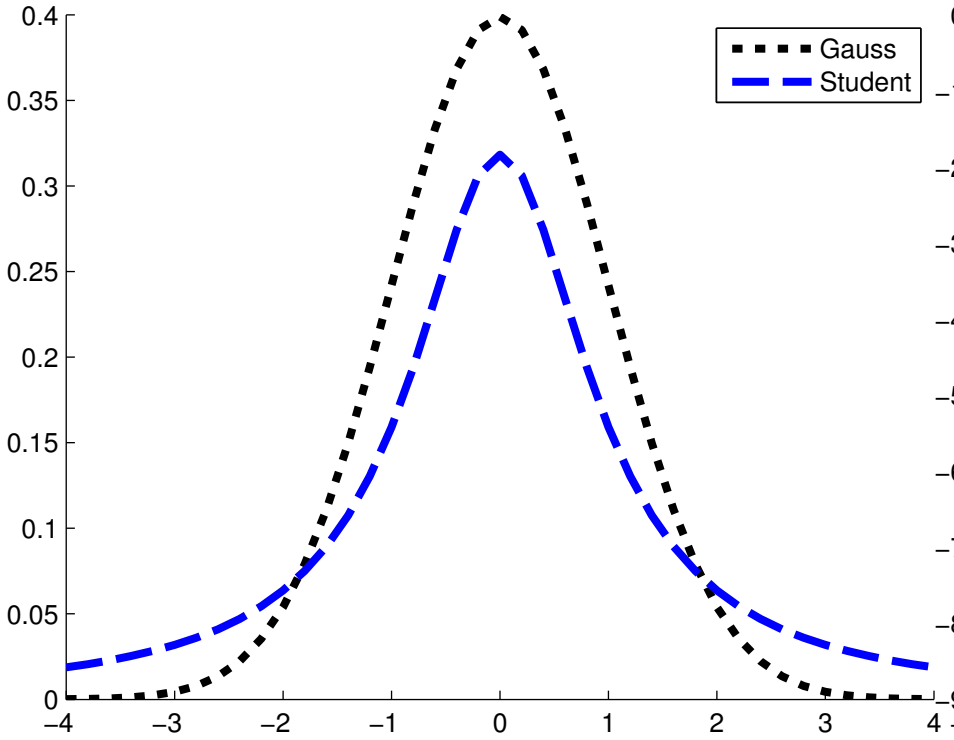
Relative to Gaussian distributions with equal variance:

- Many samples are near zero
- Occasional large-magnitude samples are far more likely
- Negative log probability density is *convex but not smooth*

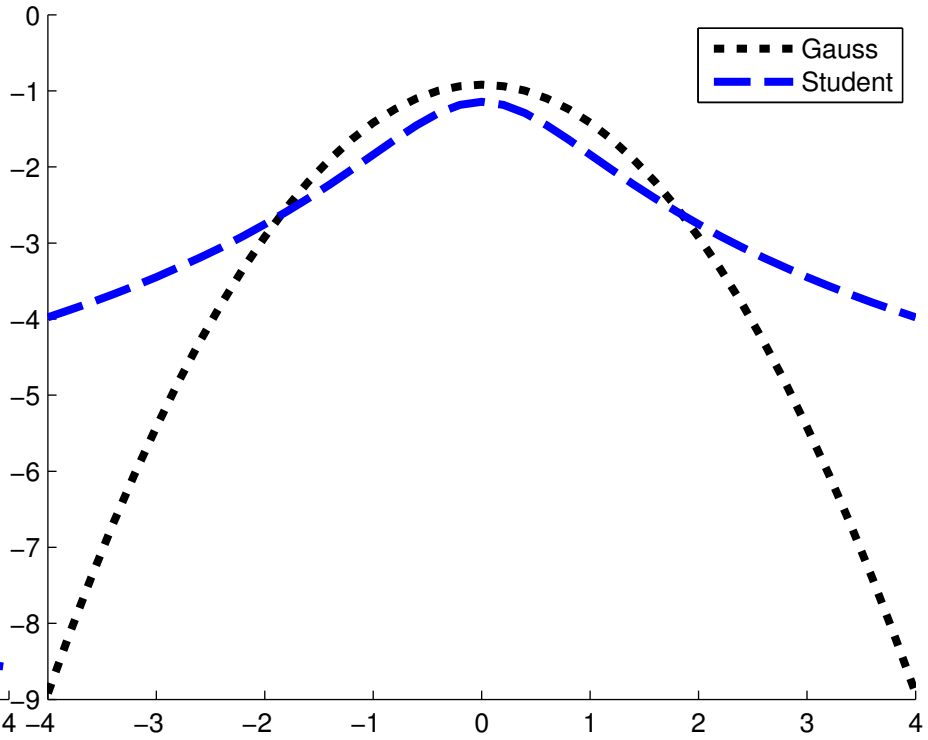
Student T Distribution

$$\mathcal{T}(x|\mu,\sigma^2,\nu) \propto \left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\left(\frac{\nu+1}{2}\right)}$$

Probability Densities



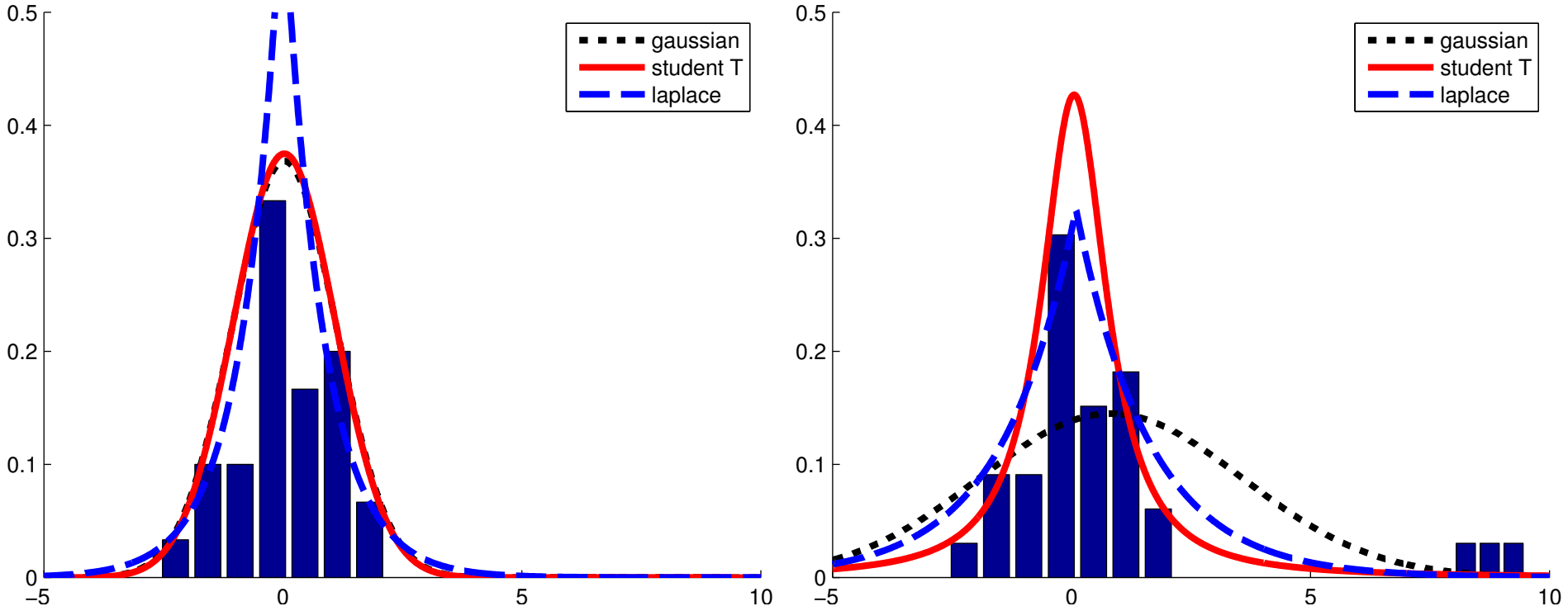
Log Probability Densities



Relative to Gaussian distributions with equal variance:

- Approaches Gaussian as DOF parameter approaches infinity
- For small DOF, large-magnitude samples are far more likely
- Negative log probability density is *smooth but not convex*

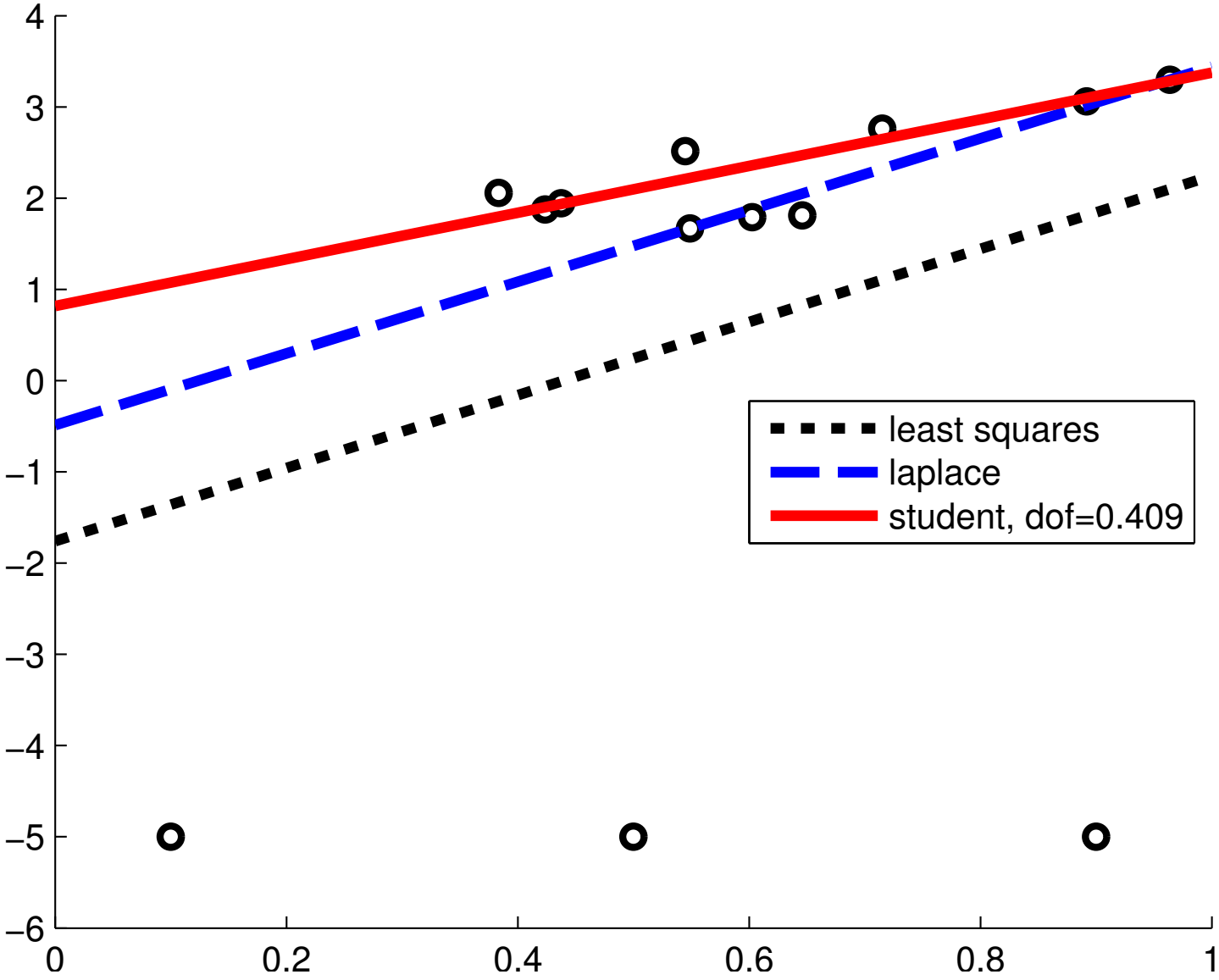
Outliers & ML Estimation



Maximum likelihood estimates of mean parameters:

- Gaussian: Sample mean of data
- Laplacian: Sample median of data
- Student T: No closed form, optimize via gradient methods

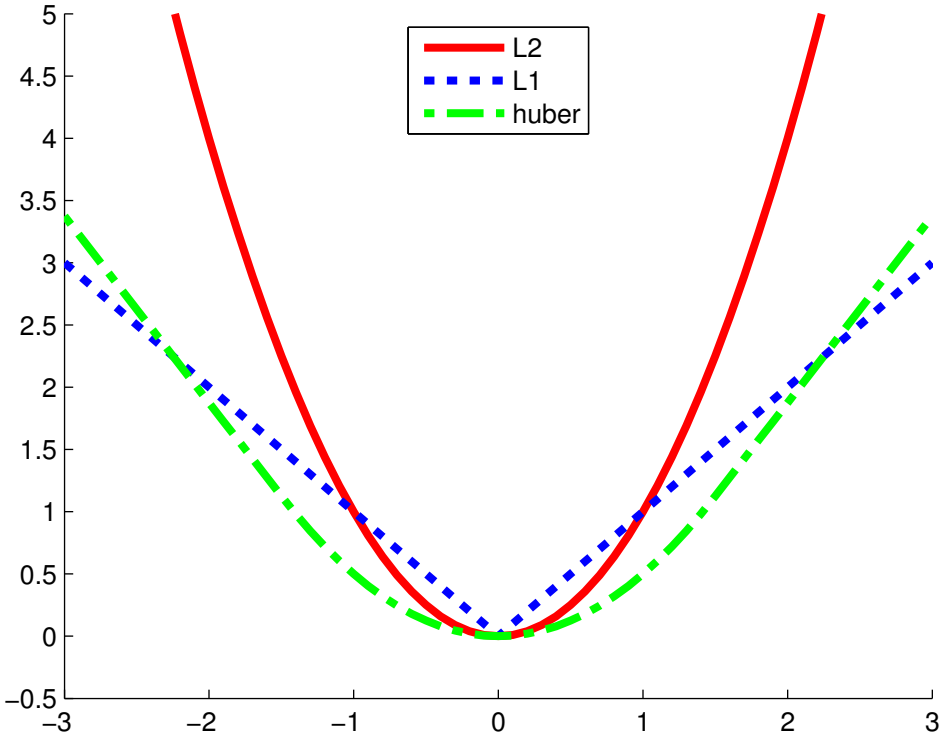
Outliers & Linear Regression



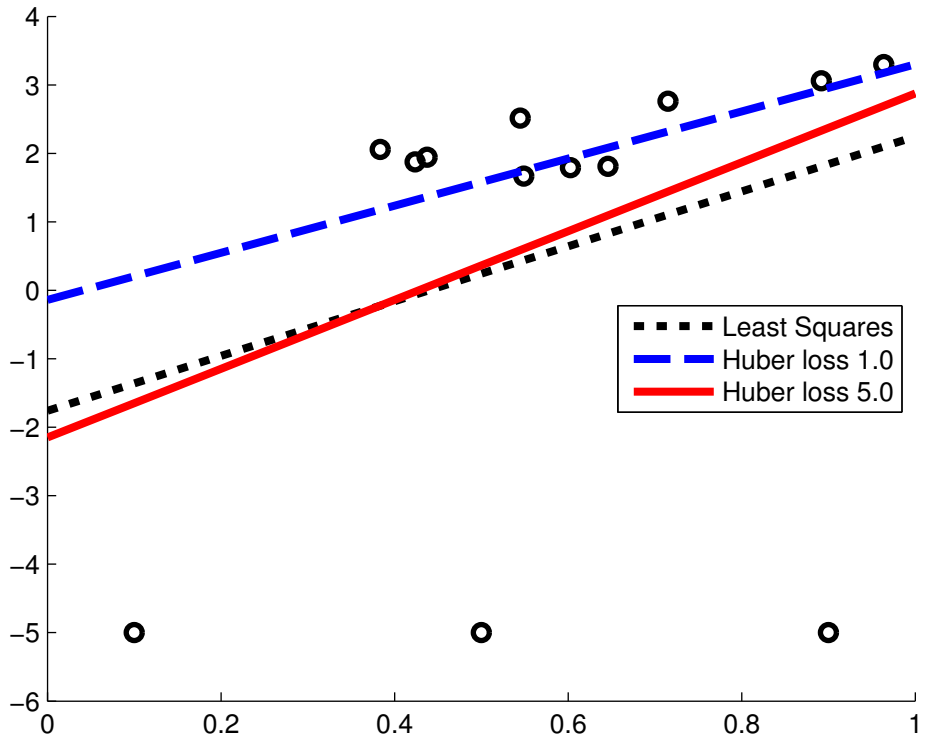
Huber Loss Function

$$L_H(r,\delta) = \begin{cases} r^2/2 & \text{if } |r| \le \delta \\ \delta|r| - \delta^2/2 & \text{if } |r| > \delta \end{cases}$$

Negative Log Probabilities



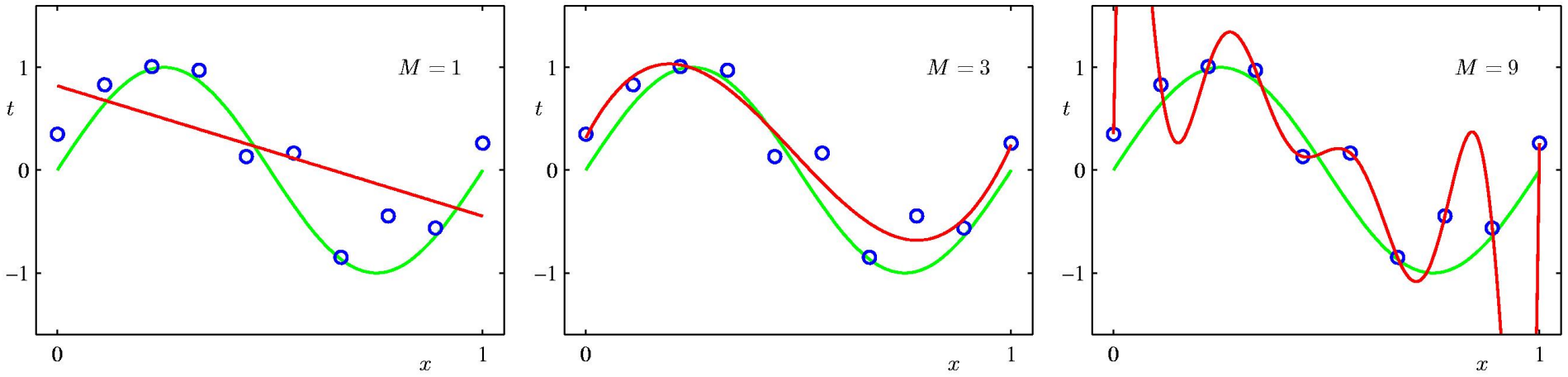
Robust Linear Regression



Relative to Gaussian distributions with equal variance:

- Behaves like Gaussian near origin (“non-outliers”)
- Behaves like Laplacian far from origin (robustness)
- Negative log probability density is *smooth and convex*

Regularization in Regression



- Basic model selection: Coefficients are ordered, and only the first *M* are non-zero
  - Classical example: polynomial regression
  - What if my features aren't easy to interpret?
- Gaussian prior (L_2 regularization): Coefficients are small
  - Computation & storage: Expensive for many features
  - Interpretability: Doesn't identify important features
- Many applications: Only some of my features are relevant, but I don't know how many or which ones

Feature Selection Models

$\phi_j(x) \in \mathbb{R}$ is some possible feature of input data

$\gamma_j = 1$ if feature j is relevant, 0 otherwise

- We would like a posterior distribution on feature inclusion:

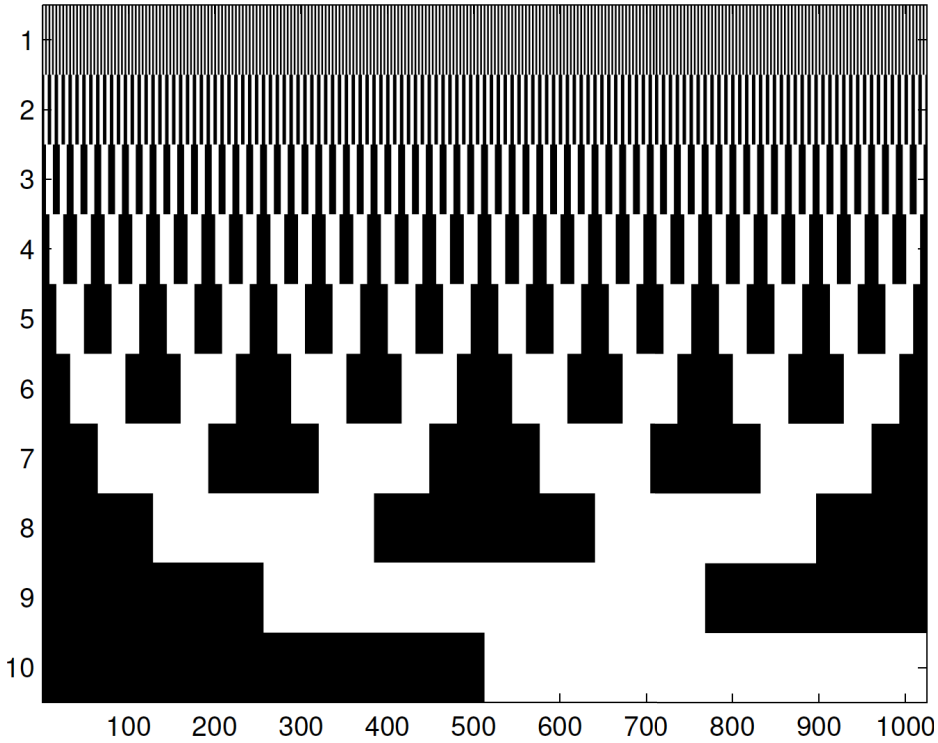
$$p(\boldsymbol{\gamma}|\mathcal{D}) = \frac{e^{-f(\boldsymbol{\gamma})}}{\sum_{\boldsymbol{\gamma}'} e^{-f(\boldsymbol{\gamma}')}} \qquad f(\boldsymbol{\gamma}) \triangleq -[\log p(\mathcal{D}|\boldsymbol{\gamma}) + \log p(\boldsymbol{\gamma})]$$

- The likelihood $p(\mathcal{D}|\boldsymbol{\gamma})$ could be any standard ML model, constrained to only depend on features for which $\gamma_j = 1$
- A common prior on the feature inclusion vector:

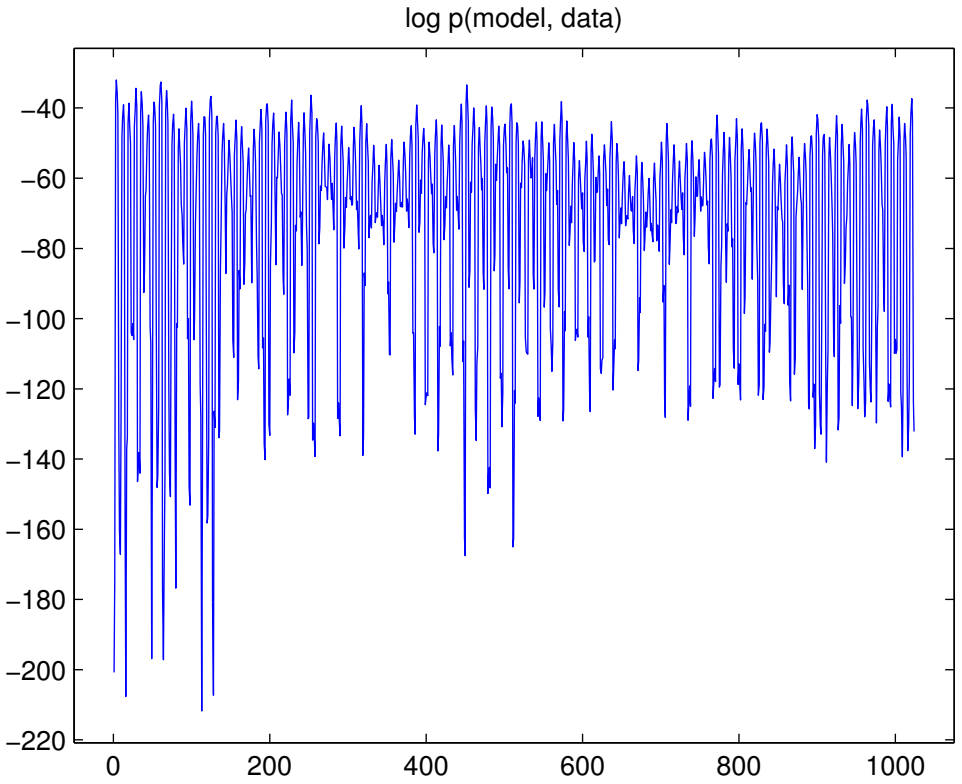
$$p(\boldsymbol{\gamma}) = \prod_{j=1}^{D} \mathrm{Ber}(\gamma_j|\pi_0) = \pi_0^{||\boldsymbol{\gamma}||_0}(1-\pi_0)^{D-||\boldsymbol{\gamma}||_0} \qquad ||\boldsymbol{\gamma}||_0 = \sum_{j=1}^{D} \gamma_j$$

$$\begin{aligned} \log p(\boldsymbol{\gamma}|\pi_0) &= ||\boldsymbol{\gamma}||_0 \log \pi_0 + (D - ||\boldsymbol{\gamma}||_0)\log(1-\pi_0) \\ &= -\lambda||\boldsymbol{\gamma}||_0 + \text{const} \end{aligned} \qquad \lambda \triangleq \log \frac{1-\pi_0}{\pi_0}$$

Feature Selection: Example



All combinations of 10 features (2^{10} in Gray code order)



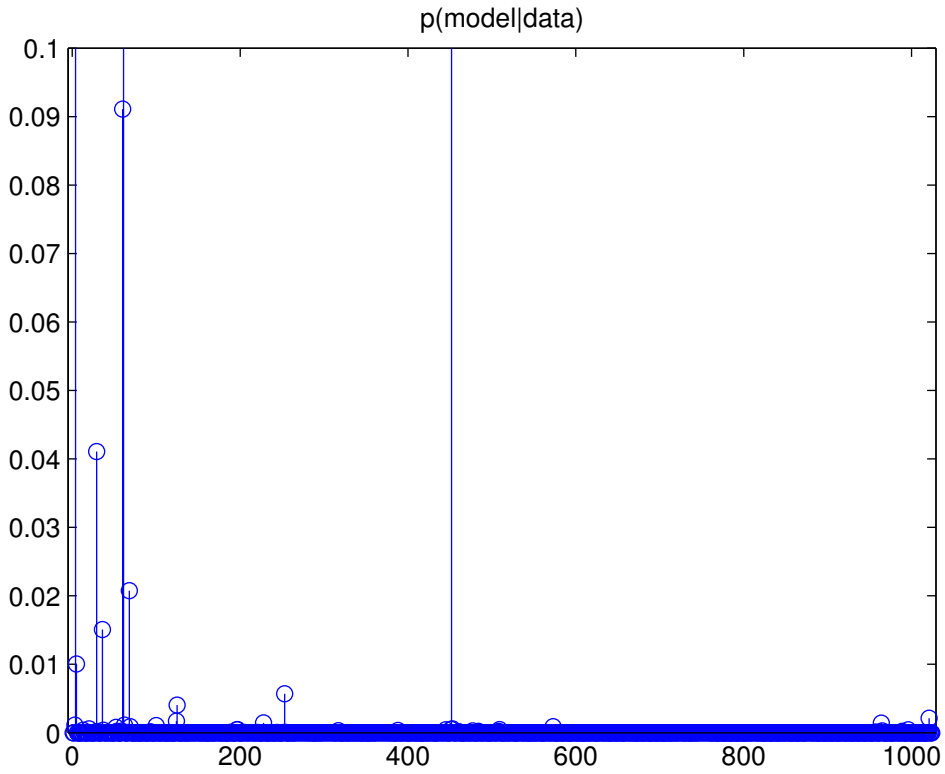
Log posterior probabilities

MAP: $\hat{\boldsymbol{\gamma}} = \operatorname{argmax} p(\boldsymbol{\gamma}|\mathcal{D}) = \operatorname{argmin} f(\boldsymbol{\gamma})$

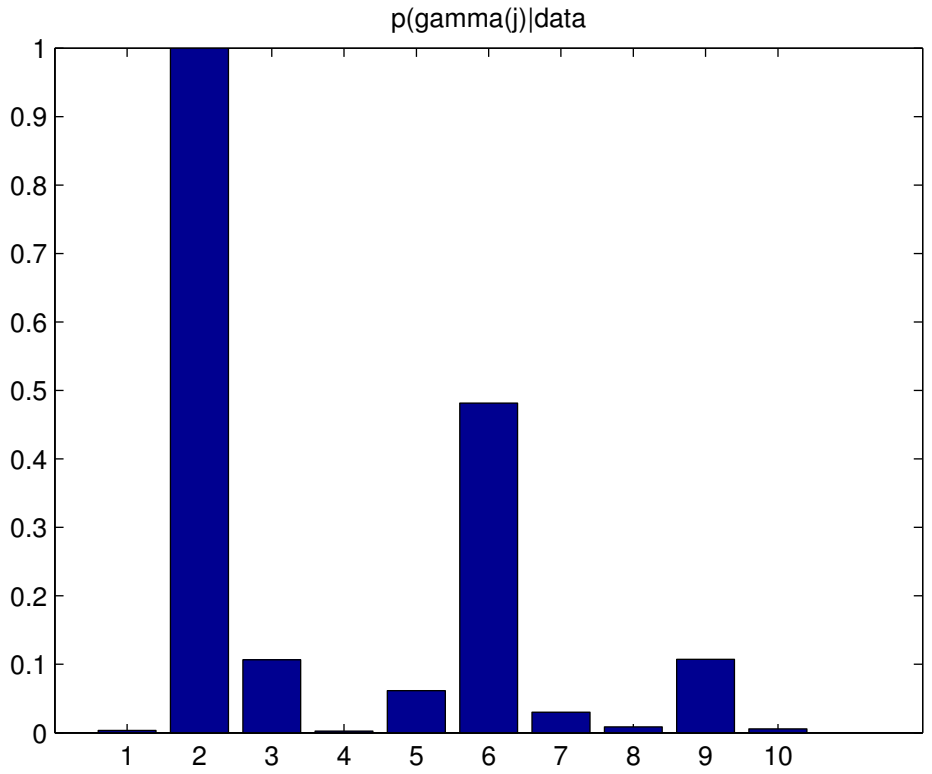
Dataset: N=10 samples based on linear regression weights

$$\mathbf{w} = (0.00, -1.67, 0.13, 0.00, 0.00, 1.19, 0.00, -0.04, 0.33, 0.00)$$

Feature Selection: Example



Posterior Probabilities



Marginal Inclusion Probabilities

Most likely models:
{2}, {2,6}, {2,6,9}, {2,3,6}…

$$\hat{\boldsymbol{\gamma}} = \{j : p(\gamma_j = 1|\mathcal{D}) > 0.5\}$$

Dataset: N=10 samples based on linear regression weights

$$\mathbf{w} = (0.00, -1.67, 0.13, 0.00, 0.00, 1.19, 0.00, -0.04, 0.33, 0.00)$$

Feature Selection for Regression

- Bernoulli prior on feature inclusion indicators:

$$p(\boldsymbol{\gamma}) = \prod_{j=1}^{D} \mathrm{Ber}(\gamma_j|\pi_0) = \pi_0^{||\boldsymbol{\gamma}||_0}(1-\pi_0)^{D-||\boldsymbol{\gamma}||_0} \qquad ||\boldsymbol{\gamma}||_0 = \sum_{j=1}^{D} \gamma_j$$

- Combine with Gaussian likelihood and weight vector prior:

$$y_i|\mathbf{x}_i, \mathbf{w}, \boldsymbol{\gamma}, \sigma^2 \sim \mathcal{N}(\sum_j \gamma_j w_j x_{ij}, \sigma^2) \qquad w_j \sim \mathcal{N}(0, \sigma_w^2)$$

- Negative log-posterior distribution:

$$f(\boldsymbol{\gamma}, \mathbf{w}) \triangleq -2\sigma^2 \log p(\boldsymbol{\gamma}, \mathbf{w}, \mathbf{y}|\mathbf{X}) = ||\mathbf{y} - \mathbf{X}(\boldsymbol{\gamma} .* \mathbf{w})||^2 + \frac{\sigma^2}{\sigma_w^2}||\mathbf{w}||^2 + \lambda||\boldsymbol{\gamma}||_0 + \mathrm{cons}$$

- Simplifying in the limit as $\sigma_w^2 \to \infty$

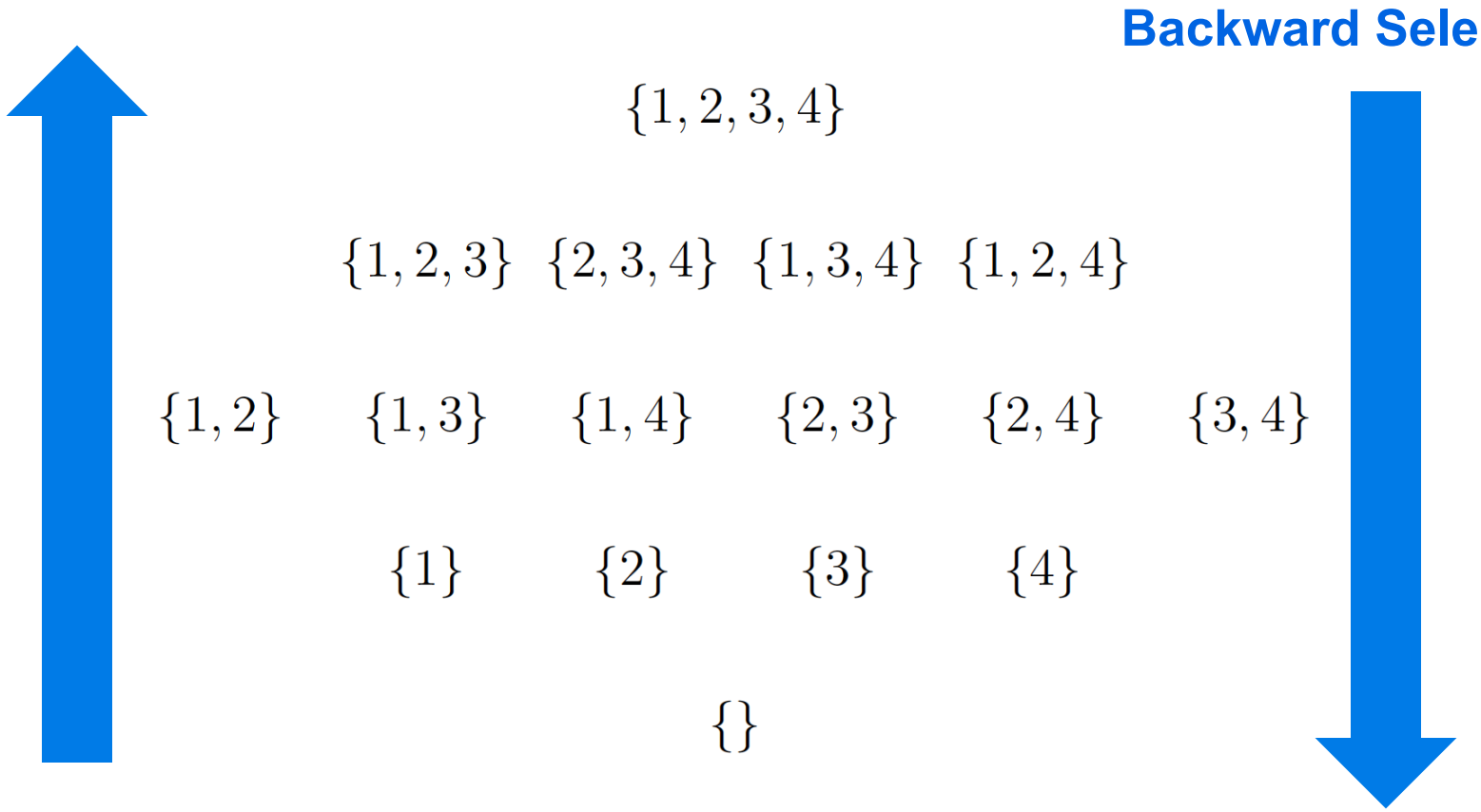
$$f(\boldsymbol{\gamma}, \mathbf{w}) = ||\mathbf{y} - \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\gamma}||_2^2 + \lambda||\boldsymbol{\gamma}||_0$$

pick out subset of feature columns

$$f(\mathbf{w}) = ||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_0$$

keep all columns but penalize non-zero weights

Greedy Deterministic Search



- Consider all possible ways of adding *(forward selection)* or removing *(backward selection)* one feature
- Add or remove the best feature, or stop if the current model is best
- Wrapper method: Can be applied to any objective. *Guarantees???*