Advanced Bayesian Learning

Regularization and Variable Selection - Lecture 1

Mattias Villani

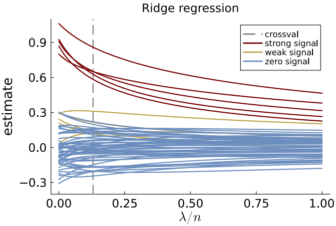
Department of Statistics
Stockholm University

Topic overview

- **Bayesian regularization priors**
  - Ridge prior
  - Lasso prior
  - Horseshoe prior
  - Dynamic shrinkage priors
- **Bayesian variable selection**
  - Spike-and-slab variable selection regression
  - Polya-Gamma augmentation for logistic regression
  - Extensions

Ridge regression (L2-regularized)

- Minimization of L2-penalized sum of squares
$$(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda\beta^T\beta$$
gives **Ridge regression**
$$\tilde{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_p\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$
- **Shrinkage** toward zero: as $\lambda \to \infty$, $\tilde{\boldsymbol{\beta}} \to 0$.
- Prevents overfitting.
- Numerical stability. Can handle $p >> n$ case.
- Estimate λ by cross-validation.



Ridge regression is an iid normal prior

- **Ridge regression** minimizes L2-penalized sum of squares

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

- Corresponds to the posterior mean under **iid normal prior**

$$\beta_j|\sigma^2 \overset{\text{iid}}{\sim} N\left(0, \frac{\sigma^2}{\lambda}\right)$$

- Note that

$$\log p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}) \propto -\frac{1}{2\sigma^2}\left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}\right]$$

  so **penalty** $=$ log **prior**.
- Gaussian has thin tails. No extreme values.
- Prior beliefs: all β_j are roughly of the same size.

Recall: Linear regression - conjugate prior

- **Joint prior** for β and σ^2

$$\beta|\sigma^2 \sim N\left(\mu_0, \sigma^2\Omega_0^{-1}\right)$$
$$\sigma^2 \sim \mathrm{Inv}-\chi^2\left(\nu_0, \sigma_0^2\right)$$

- **Posterior**

$$\beta|\sigma^2, \mathsf{y} \sim N\left(\mu_n, \sigma^2\Omega_n^{-1}\right)$$
$$\sigma^2|\mathsf{y} \sim \mathrm{Inv}{-}\chi^2\left(\nu_n, \sigma_n^2\right)$$

$$\mu_n = \left(\mathsf{X}^\top\mathsf{X} + \Omega_0\right)^{-1}\left(\mathsf{X}^\top\mathsf{X}\hat{\beta} + \Omega_0\mu_0\right)$$
$$\Omega_n = \mathsf{X}^\top\mathsf{X} + \Omega_0$$
$$\nu_n = \nu_0 + n$$
$$\sigma_n^2 = \left(\nu_0\sigma_0^2 + \mathsf{y}^\top\mathsf{y} + \mu_0^\top\Omega_0\mu_0 - \mu_n^\top\Omega_n\mu_n\right)/\nu_n$$

Direct sampling L2-regularization parameter

- Cross-validation used to determine degree of smoothness, λ.
- Bayesian: λ is unknown $\Rightarrow$ put a prior for λ!
- The joint posterior of β, σ^2 and λ is $(\Omega_0(\lambda) = \lambda I)$

$$
\begin{aligned}
\beta|\sigma^2, \lambda, \mathrm{y}, \mathrm{X} &\sim N\left(\mu_n, \Omega_n^{-1}\right) \\
\sigma^2|\lambda, \mathrm{y}, \mathrm{X} &\sim \mathrm{Inv}-\chi^2\left(\nu_n, \sigma_n^2\right) \\
p(\lambda|\mathrm{y}, \mathrm{X}) &\propto \sqrt{\frac{|\Omega_0(\lambda)|}{\left|\boldsymbol{X}^\top\boldsymbol{X} + \Omega_0(\lambda)\right|}}\left(\frac{\nu_n\sigma_n^2(\lambda)}{2}\right)^{-\nu_n/2} \cdot p(\lambda)
\end{aligned}
$$

- This is the conditional-marginal decomposition

$$
p(\beta, \sigma^2, \lambda|\mathrm{y}, \mathrm{X}) = p(\beta|\sigma^2, \lambda, \mathrm{y}, \mathrm{X})p(\sigma^2|\lambda, \mathrm{y}, \mathrm{X})p(\lambda|\mathrm{y}, \mathrm{X})
$$

Gibbs sampling for L2-regularized regression

- Prior:

$$\boldsymbol{\beta}|\sigma^2, \lambda \sim N\left(0, \frac{\sigma^2}{\lambda} I_p\right)$$
$$\sigma^2 \sim \text{Inv}-\chi^2\left(\nu_0, \sigma_0^2\right)$$
$$\lambda^{-1} \sim \text{Inv}-\chi^2\left(\omega_0, \psi_0^2\right).$$

- By Bayes' theorem

$$p\left(\lambda|\boldsymbol{\beta}, \sigma^2, \mathrm{y}\right) \propto p\left(\mathrm{y}|\boldsymbol{\beta}, \sigma^2, \lambda\right) p\left(\lambda|\boldsymbol{\beta}, \sigma^2\right)$$

- $p\left(\mathrm{y}|\boldsymbol{\beta}, \sigma^2, \lambda\right)$ does not depend on λ once we condition on $\boldsymbol{\beta}$:

$$p\left(\lambda|\boldsymbol{\beta}, \sigma^2, \mathrm{y}\right) \propto p\left(\lambda|\boldsymbol{\beta}, \sigma^2\right)$$

- So using Bayes' theorem once more

$$p\left(\lambda|\boldsymbol{\beta}, \sigma^2\right) \propto p\left(\boldsymbol{\beta}|\sigma^2, \lambda\right) p\left(\lambda\right)$$

- In conditional posterior for λ, the $\beta_1, \ldots, \beta_p$ act like "data".

Gibbs sampling L2-regularized regression $\psi^2 = \lambda^{-1}$

Gibbs sampling linear regression - L2 regularization prior

The posterior for the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon,\ \varepsilon \sim N(\mathbf{0}, \sigma^2 I_n), \tag{12.16}$$

with hierarchical L2 regularization prior

$$\begin{aligned}
\boldsymbol{\beta}|\sigma^2, \psi^2 &\sim N(\mathbf{0}, \sigma^2\psi^2 I_p) \\
\sigma^2 &\sim \text{Inv}-\chi^2(\nu_0, \tau_0^2) \\
\psi^2 &\sim \text{Inv}-\chi^2(\omega_0, \psi_0^2).
\end{aligned}$$

can be sampled by a two-block Gibbs sampler:

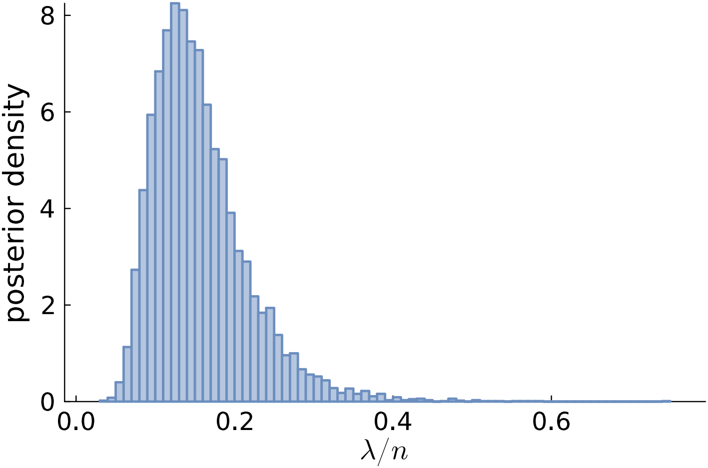
$$\begin{aligned}
\text{Block1}:\ \boldsymbol{\beta}|\sigma^2, \psi^2, \mathbf{y} &\sim N\left(\hat{\boldsymbol{\beta}}_{L_2}, \sigma^2(\mathbf{X}^\top\mathbf{X} + \psi^{-2}I_p)^{-1}\right) \\
\sigma^2|\psi^2, \mathbf{y} &\sim \text{Inv}-\chi^2(\tau_n^2, \nu_n) \\
\text{Block2}:\ \psi^2|\boldsymbol{\beta}, \sigma^2, \mathbf{y} &\sim \text{Inv}-\chi^2(\omega_n, \psi_n^2),
\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_{L_2}$ is the ridge estimator

$$\hat{\boldsymbol{\beta}}_{L_2} = \left(\mathbf{X}^\top\mathbf{X} + \psi^{-2}I_p\right)^{-1}\mathbf{X}^\top\mathbf{y} = \left(\mathbf{X}^\top\mathbf{X} + \lambda I_p\right)^{-1}\mathbf{X}^\top\mathbf{y}.$$

The hyperparameters ν_n and τ_n^2 are given in Figure 5.3.
Finally, $\omega_n = \omega_0 + p$ and $\psi_n^2 = \left(\sum_{i=1}^p (\beta_i/\sigma)^2 + \omega_0\psi_0^2\right)/\omega_n$.

Marginal posterior of λ



Regularization prior - Lasso

- **Lasso** is equivalent to posterior mode under Laplace prior

$$\beta_i|\sigma^2 \overset{\text{iid}}{\sim} \text{Laplace}\left(0, \frac{\sigma^2}{2\lambda}\right)$$

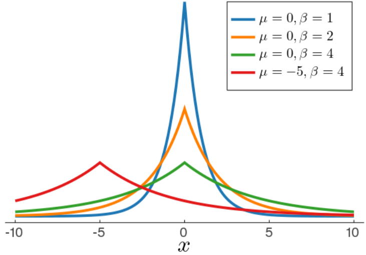
Laplace distribution

$X \sim \text{Laplace}(\mu, \beta)$ for $X \in \mathbb{R}$.

$$p(x) = \frac{1}{2\beta}\exp\left(-\frac{|x-\mu|}{\beta}\right)$$

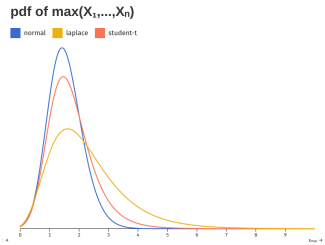
$$\mathbb{E}(X) = \mu$$

$$\mathbb{V}(X) = 2\beta^2$$

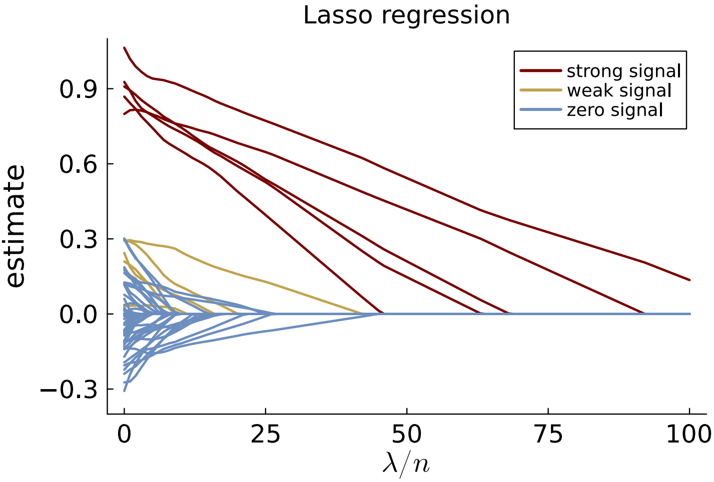


- Laplace distribution has heavier tails than normal.
- **Laplace**: many β_i close to zero, but some β_i rather large.

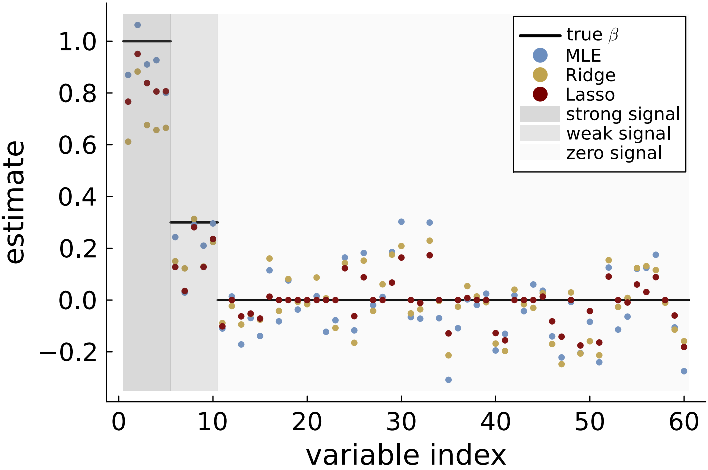
A tale of tails



Lasso/Laplace prior



Ridge vs Lasso shrinkage



Horseshoe prior

- Normal and Laplace - only one global shrinkage parameter λ.
- **Global-Local shrinkage**: global + local shrinkage for each β_j.
- **Horseshoe prior**:

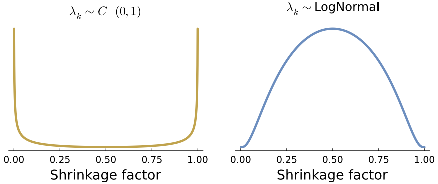
$$\beta_j|\lambda_j^2, \tau^2 \sim N\left(0, \tau^2\lambda_j^2\right)$$

$$\lambda_j \overset{\text{iid}}{\sim} C^+(0,1)$$

$$\tau \sim C^+(0,1)$$

- **Shrinkage factor** c_j (orthogonal covariates)

$$\tilde{\beta}_j = (1-c_j)\hat{\beta}_j, \qquad c_j = \frac{1}{1+(n/\sigma^2)\tau^2\lambda_j^2}$$



Gibbs sampling for regression with horseshoe prior

- $X \sim C^{+}(0,1)$ can be generated by continuous mixture:

$$Y \sim \mathrm{Inv-}\chi^2(1,2)$$
$$X^2|Y \sim \mathrm{Inv-}\chi^2(1,2/Y)$$

- **Horseshoe prior** in mixture formulation:

$$\boldsymbol{\beta}|\lambda_1,\ldots,\lambda_p,\tau^2,\sigma^2\boldsymbol{\Lambda} \sim N\left(0,\sigma^2\tau^2\boldsymbol{\Lambda}\right)$$
$$\lambda_j^2|\nu_j \overset{\text{inde}}{\sim} \mathrm{Inv-}\chi^2(1,2/\nu_j)$$
$$\tau^2|\xi \sim \mathrm{Inv-}\chi^2(1,2/\xi)$$
$$\nu_1,\ldots,\nu_p,\xi \overset{\text{iid}}{\sim} \mathrm{Inv-}\chi^2(1,2)$$

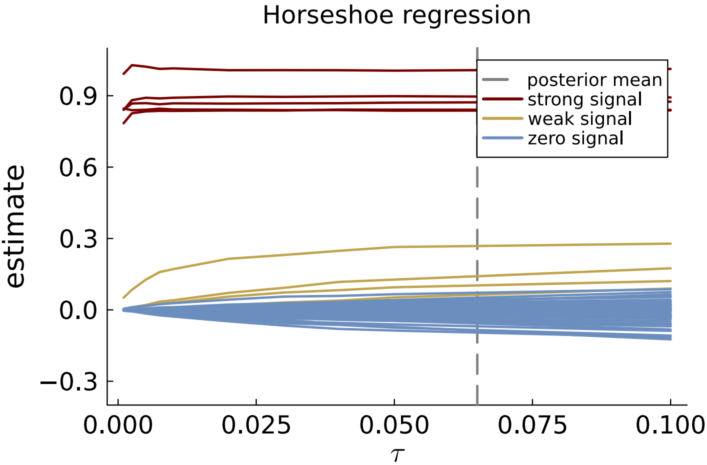
where $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_1^2,\ldots,\lambda_p^2)$.

Gibbs sampling for regression with horseshoe prior

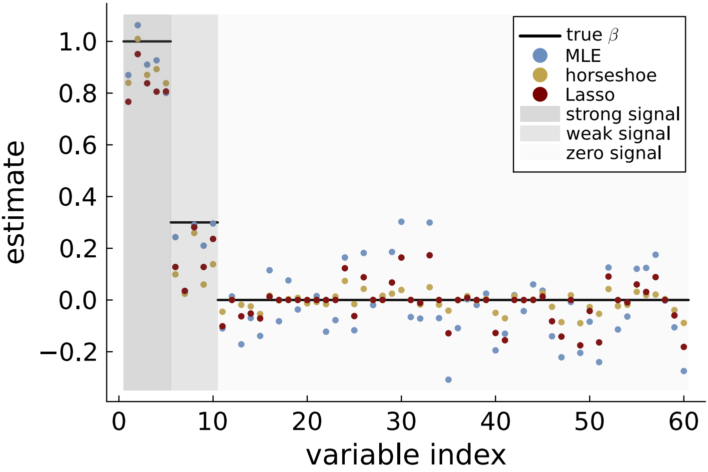
- Gibbs sampler

$$
\begin{aligned}
\boldsymbol{\beta}, \sigma | \boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{X} &\sim \text{Linear regression with } \boldsymbol{\Omega}_0^{-1} = \tau^2 \boldsymbol{\Lambda} \\
\nu_j | \lambda_j &\overset{\text{iid}}{\sim} \text{Inv}-\chi^2(2, 1 + 1/\lambda_j^2) \\
\lambda_j^2 | \nu_j, \tau, \boldsymbol{\beta}, \sigma &\sim \text{Inv}-\chi^2\left(2, \frac{1}{\nu_j} + \frac{1}{2}\left(\frac{\beta_j}{\sigma\tau}\right)^2\right) \\
\xi | \tau &\overset{\text{iid}}{\sim} \text{Inv}-\chi^2(2, 1 + 1/\tau^2), \\
\tau^2 | \xi, \lambda_1, \ldots, \lambda_p, \boldsymbol{\beta}, \sigma &\sim \text{Inv}-\chi^2\left(p + 1, \frac{\frac{2}{\xi} + \sum_{j=1}^p \left(\frac{\beta_j}{\sigma\lambda_j}\right)^2}{p+1}\right)
\end{aligned}
$$

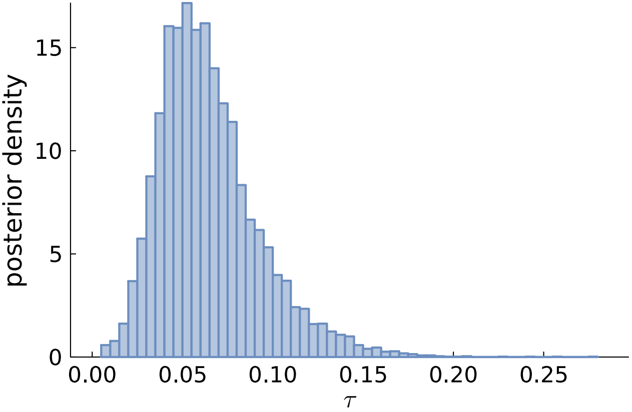
Horseshoe prior on simulated data



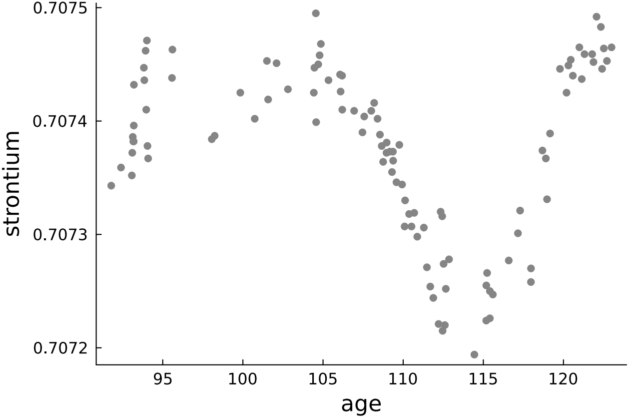
Horseshoe prior on simulated data



Horseshoe prior on simulated data



Spline regression - fossil data case study



Polynomial regression

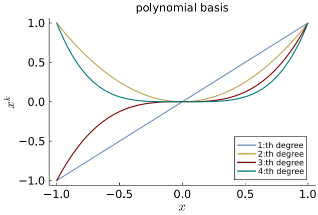
- **Polynomial regression**

$$f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_k x_i^k, \quad \text{for } i = 1, \ldots, n.$$

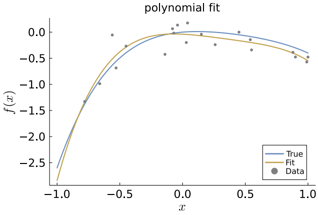
$$\mathrm{y} = \boldsymbol{X}\beta + \varepsilon,$$

$$\boldsymbol{x}_i = (1, x_i, x_i^2, \ldots, x_i^k)^\top$$

- Still linear in β and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. Bayes unchanged.

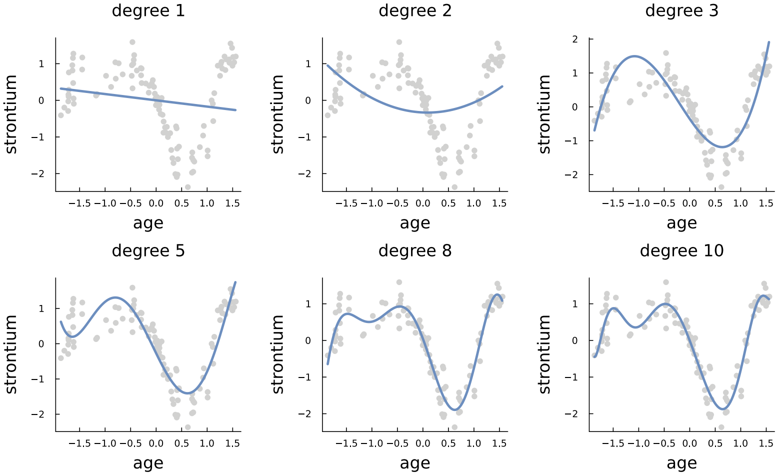


1.00	-1.00	1.00	-1.00	1.00
1.00	-0.90	0.81	-0.73	0.66
1.00	-0.80	0.64	-0.51	0.41
1.00	-0.70	0.49	-0.34	0.24
1.00	-0.60	0.36	-0.22	0.13
1.00	-0.50	0.25	-0.12	0.06
1.00	-0.40	0.16	-0.06	0.03
1.00	-0.30	0.09	-0.03	0.01
1.00	-0.20	0.04	-0.01	0.00
1.00	-0.10	0.01	-0.00	0.00
1.00	0.00	0.00	0.00	0.00
1.00	0.10	0.01	0.00	0.00
1.00	0.20	0.04	0.01	0.00
1.00	0.30	0.09	0.03	0.01
1.00	0.40	0.16	0.06	0.03
1.00	0.50	0.25	0.12	0.06
1.00	0.60	0.36	0.22	0.13
1.00	0.70	0.49	0.34	0.24
1.00	0.80	0.64	0.51	0.41
1.00	0.90	0.81	0.73	0.66
1.00	1.00	1.00	1.00	1.00



- Polynomials are global basis functions. Local basis preferred.

Polynomial regression - fossil data



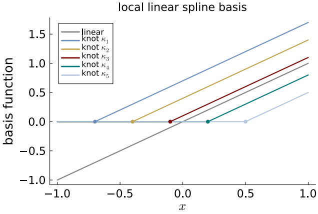
Spline regression - local linear basis

- **Truncated linear splines** with **knot locations** $\kappa_1, \ldots, \kappa_m$:

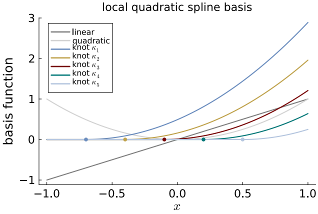
$$b_j(x) = \begin{cases} |x - \kappa_j|^p & \text{if } x > \kappa_j \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{y} = \boldsymbol{X}\beta + \varepsilon,$$

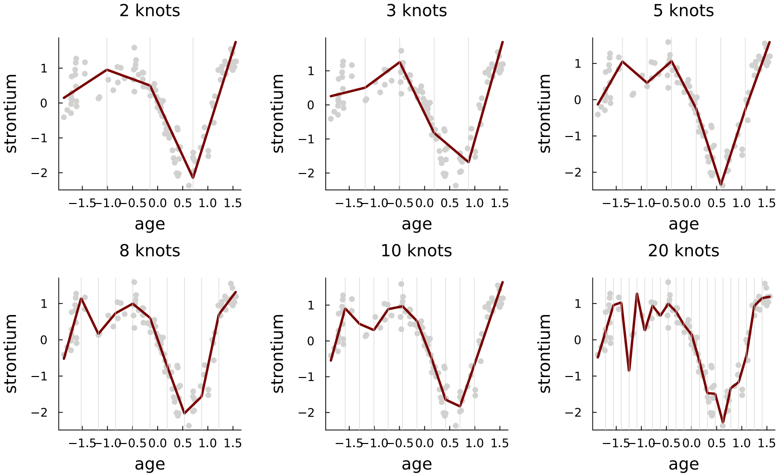
$$\boldsymbol{x}_i = (1, x_i, b_1(x_i), \ldots, b_m(x_i))^\top$$



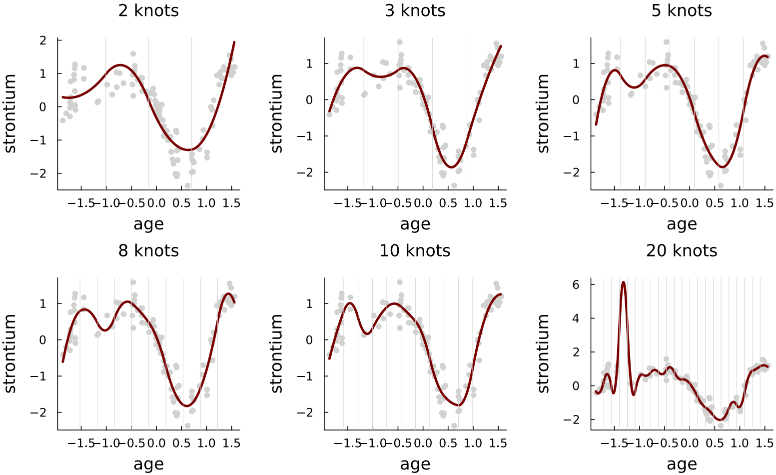
1.00	-1.00	-0.00	-0.00	-0.00
1.00	-0.90	-0.00	-0.00	-0.00
1.00	-0.80	-0.00	-0.00	-0.00
1.00	-0.70	-0.00	-0.00	-0.00
1.00	-0.60	-0.00	-0.00	-0.00
1.00	-0.50	0.00	-0.00	-0.00
1.00	-0.40	0.10	-0.00	-0.00
1.00	-0.30	0.20	-0.00	-0.00
1.00	-0.20	0.30	-0.00	-0.00
1.00	-0.10	0.40	-0.00	-0.00
1.00	0.00	0.50	0.00	-0.00
1.00	0.10	0.60	0.10	-0.00
1.00	0.20	0.70	0.20	-0.00
1.00	0.30	0.80	0.30	-0.00
1.00	0.40	0.90	0.40	-0.00
1.00	0.50	1.00	0.50	0.00
1.00	0.60	1.10	0.60	0.10
1.00	0.70	1.20	0.70	0.20
1.00	0.80	1.30	0.80	0.30
1.00	0.90	1.40	0.90	0.40
1.00	1.00	1.50	1.00	0.50



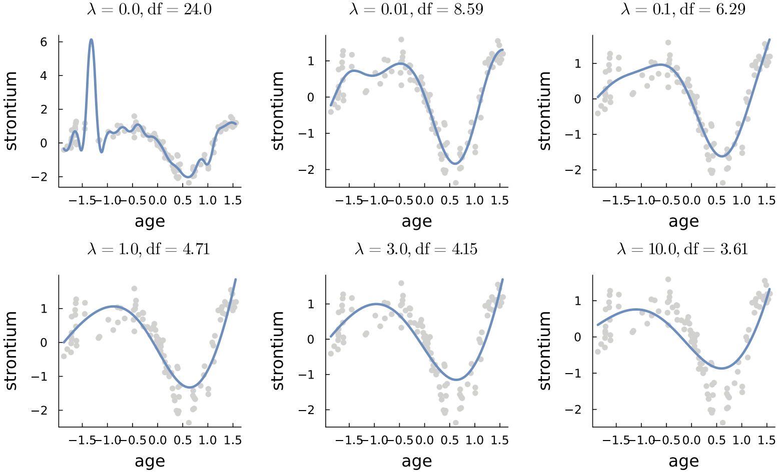
Spline regression - local linear basis



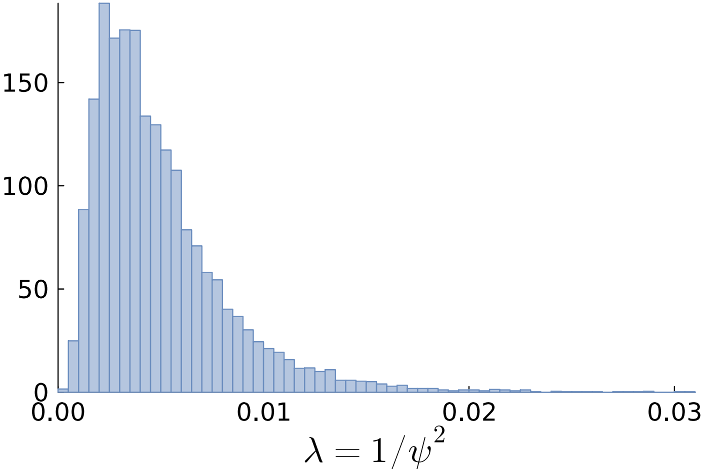
Spline regression - local quadratic basis



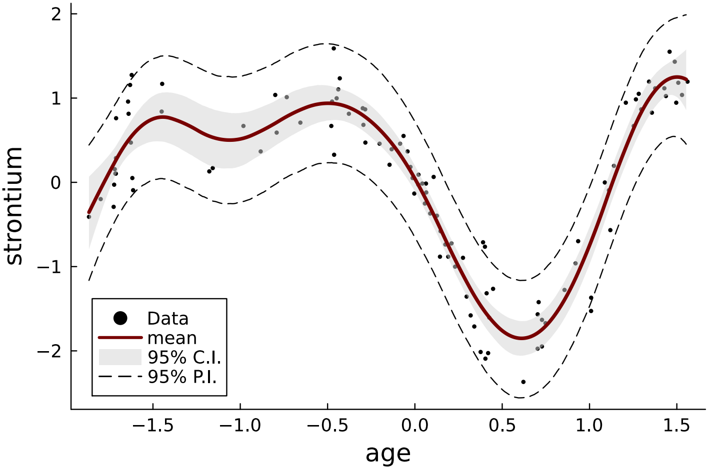
Spline regression - L2-regularization



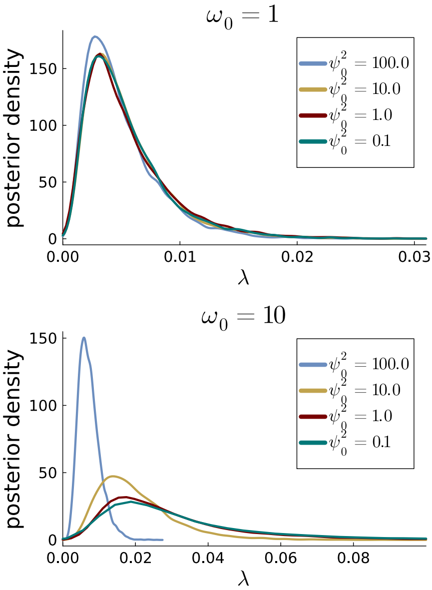
Spline regression - posterior for λ



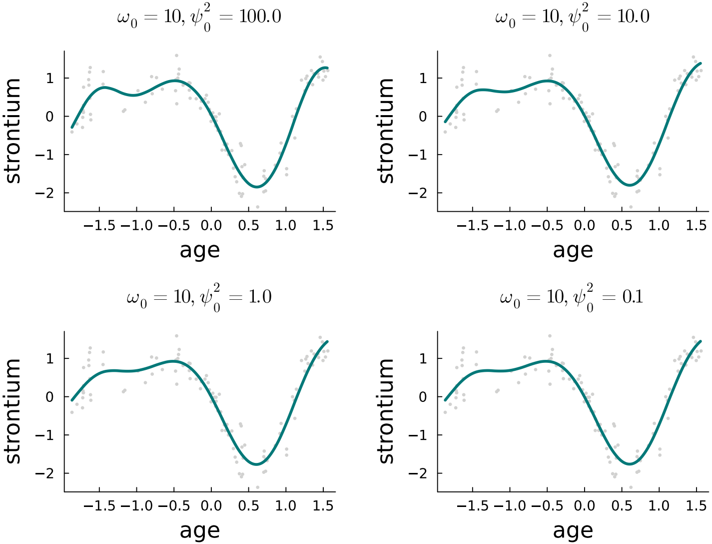
Posterior and predictive distribution



Prior sensitivity $\lambda = 1/\psi^2$ for $\psi^2 \sim \text{Inv}-\chi^2(\omega_0, \psi_0^2)$



Prior sensitivity fit $\psi^2 \sim \text{Inv}-\chi^2(\omega_0, \psi_0^2)$



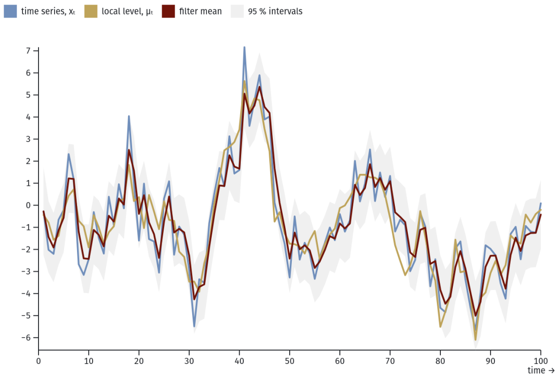
Regularization in state-space models

- **Local level model** (state-space) for time series

$$
\begin{aligned}
x_t &= \mu_t + \varepsilon_t, \qquad & \varepsilon_t &\sim N(0, \sigma_\varepsilon^2) \\
\mu_t &= \mu_{t-1} + \nu_t, \qquad & \nu_t &\sim N(0, \sigma_\nu^2)
\end{aligned}
$$

- **Innovation variance** $\sigma_\nu^2 \Rightarrow$ how fast the mean evolves.
- Same normal $N(0, \sigma_\nu^2)$ for all ν_t. Compare Ridge regression.
- Restrictive parameter evolution. Can't get all of this:
  1. $\nu_t \approx 0$ for some t (parameters stand still)
  2. large ν_t for some t (jumps)
  3. persistent periods of rapid changes

Local level model with Gaussian innovations



Dynamic horseshoe process prior

- **Horseshoe prior** for time series

$$
\begin{aligned}
\mu_t &= \mu_{t-1} + \nu_t, \qquad \nu_t \sim N(0, \tau^2\lambda_j^2)\\
\lambda_t &\overset{\text{iid}}{\sim} C^+(0,1)\\
\tau &\sim C^+(0,1)
\end{aligned}
$$

- This gives us Property 1 and 2 above. 😄
- Local variances λ_t^2 are independent. No Property 3. 🥺
- **Dynamic horseshoe process** [1]

$$
\begin{aligned}
\mu_t &= \mu_{t-1} + \nu_t, \qquad \nu_t \sim N(0, \tau^2 \exp(h_t))\\
h_t &= \phi h_{t-1} + \eta_t, \qquad \eta_t \sim Z(1/2, 1/2, 0, 1)\\
\tau &\sim C^+(0,1)
\end{aligned}
$$

- The horseshoe prior is the special case with $\phi = 0$

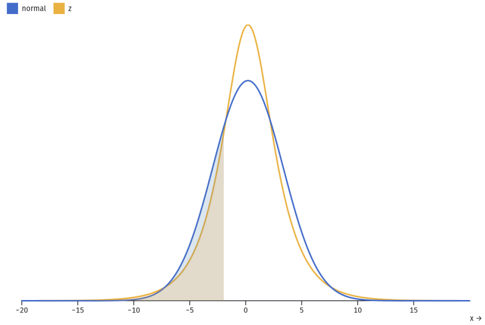
$$
\eta_t \sim Z(1/2, 1/2, 0, 1) \Longleftrightarrow \lambda_t = \exp(\eta_t/2) \sim C^+(0,1)
$$

Z-distribution

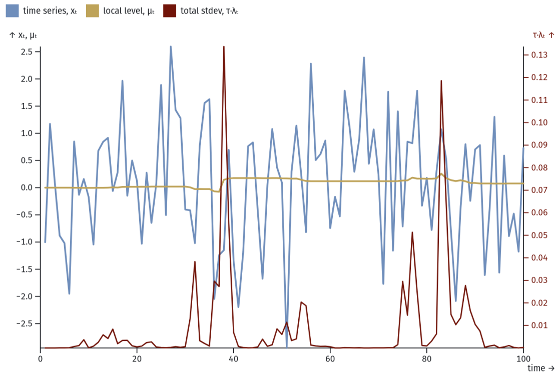
- Also called **Logistic-Beta distribution** since

$$X \sim \mathrm{Beta}(\alpha, \beta) \implies \log\left(\frac{X}{1-X}\right) \sim Z(\alpha, \beta, 0, 1)$$

- The $Z(\alpha, \beta, 0, 1)$ distribution is heavy tailed.
- Linearly decaying log density.



Local level with dynamic shrinkage process



D. R. Kowal, D. S. Matteson, and D. Ruppert, "Dynamic shrinkage processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 81, no. 4, pp. 781–804, 2019.