

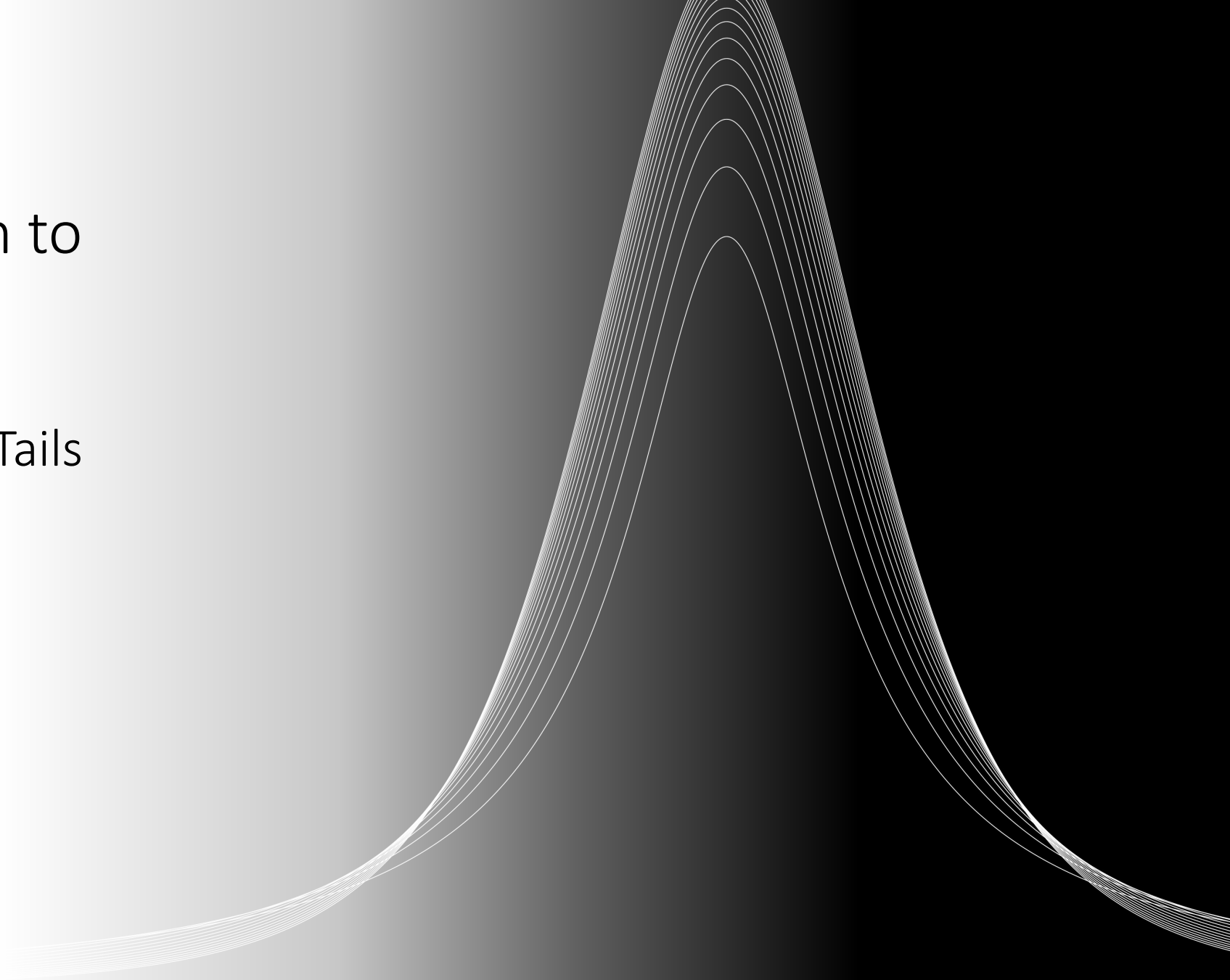


A Gentle Introduction to Bayesian Estimation

Day 4: Priors: Cautionary Tails
and Possibilities

Sara van Erp

s.j.vanerp@uu.nl



Recap day 3

Part 1: Software and algorithms

- Different ways to get the posterior
- What is going on (conceptually) under the hood?
- What should you, as user, be aware of?

Part 2: Predictive checks

- Posterior predictive checks: how can we check our model?
- Prior predictive checks

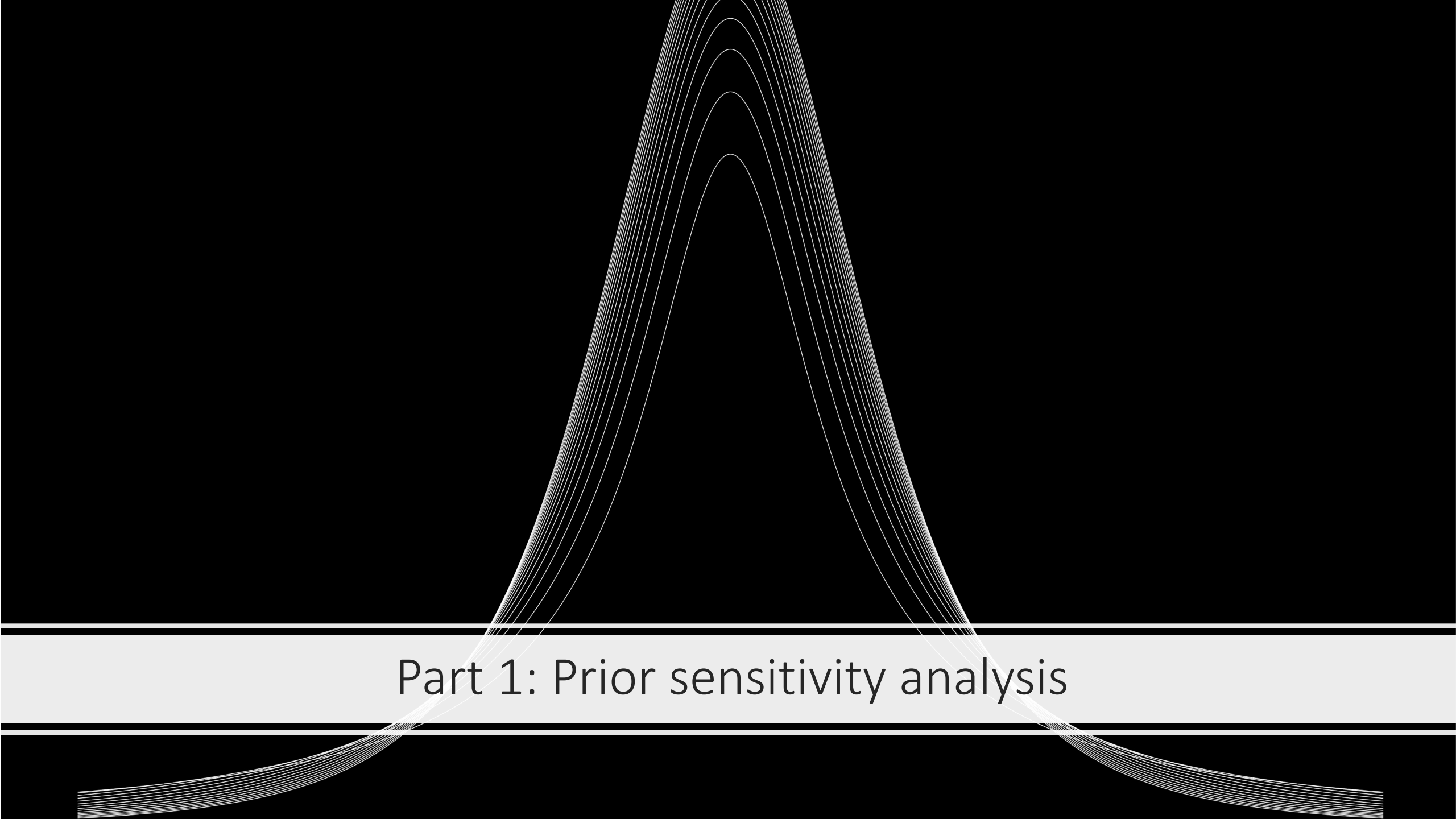
Today

Part 1: Prior sensitivity analysis

- Recap: What is a prior?
- When is a prior influential?
- How to perform a prior sensitivity analysis

Part 2: Shrinkage priors

- Basic idea behind penalization
- Different shrinkage priors = different behaviors
- Practical considerations
- Advanced applications



Part 1: Prior sensitivity analysis

Recap: The prior distribution

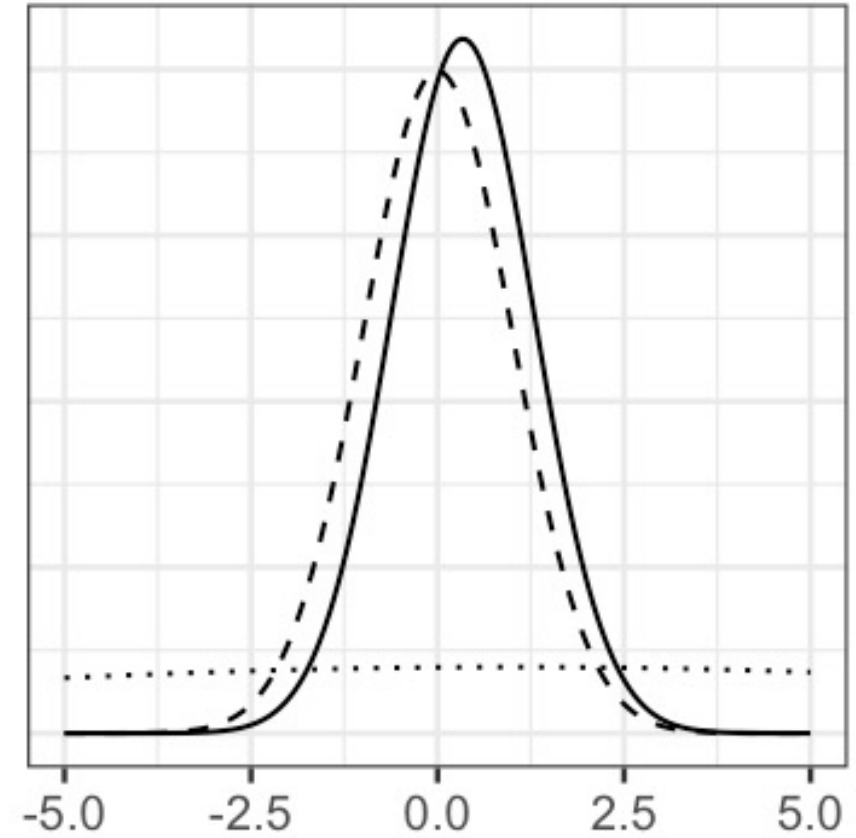
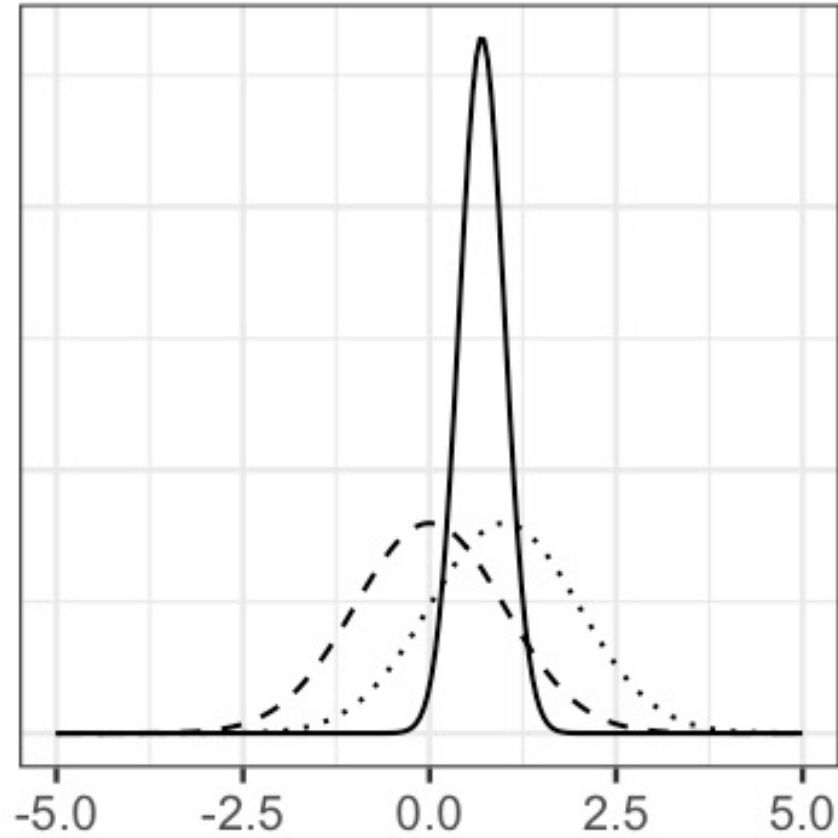
- A probability distribution
- Represents prior knowledge
- Based on previous studies, experts, data (EB), general knowledge or to serve a specific purpose (e.g., shrinkage priors)
- Varies in informativeness
- Needs to be specified for every parameter in the model

When is a prior influential?

- Highly problem-specific!



One prior, different scales data



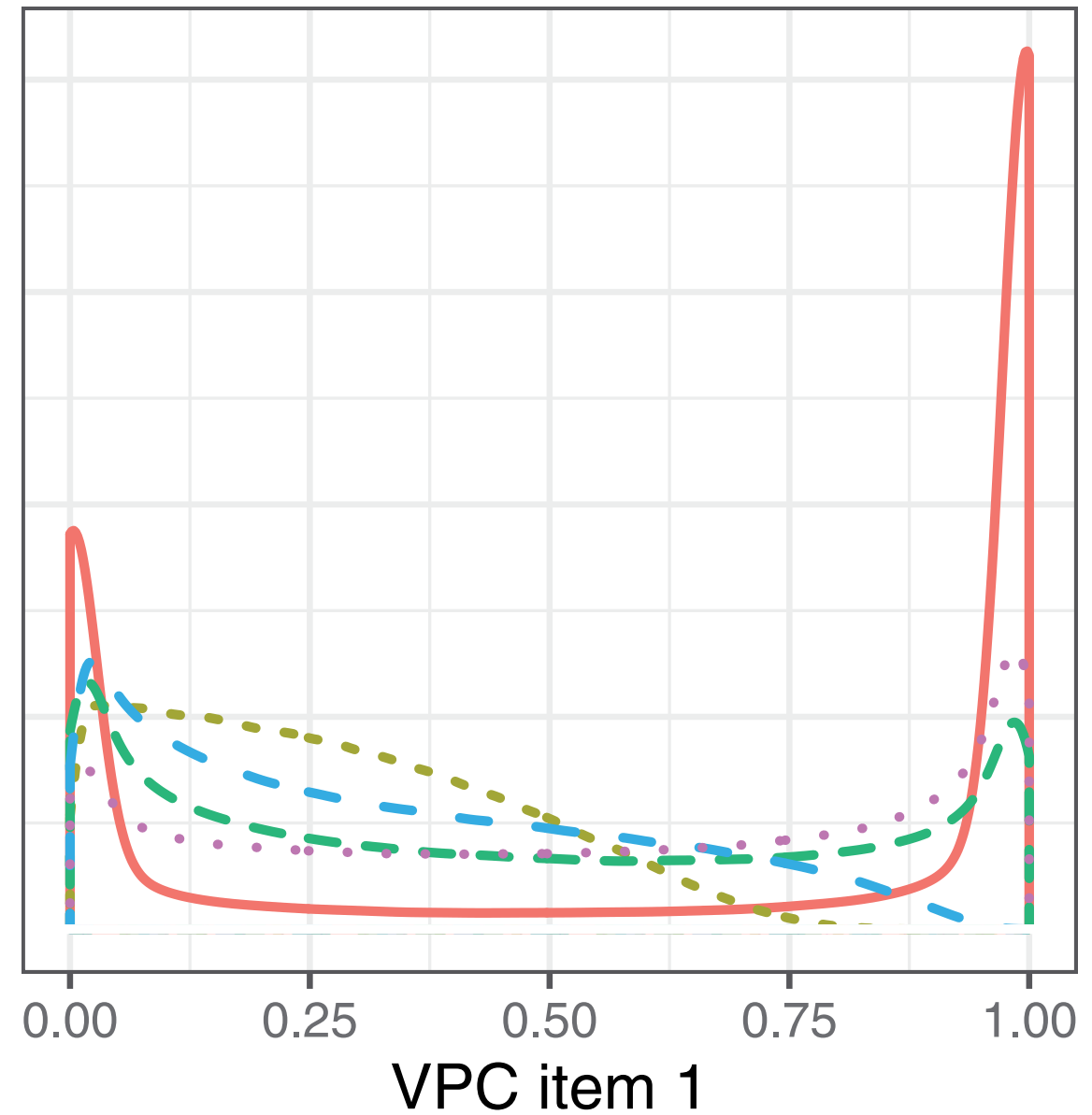
..... Likelihood — Posterior - - Prior

When is a prior influential?

- Highly problem-specific!
- Higher-level variances (multilevel, SEM) can be especially sensitive
- Implied priors on functions of parameters can prove influential



Implied priors variance partition coefficient (VPC)



From: van Erp & Browne (2021)



When is a prior influential?

- Highly problem-specific!
- Higher-level variances (multilevel, SEM) can be especially sensitive
- Implied priors on functions of parameters can prove influential

Conclusion

- Understand your prior as well as possible before the analysis (visualizations, prior predictive checks)
- Conduct a prior sensitivity analysis afterwards to check your understanding



Prior sensitivity analysis

Basic idea

Rerunning the analysis with different priors, although automatic procedures exist.

Ideal situation

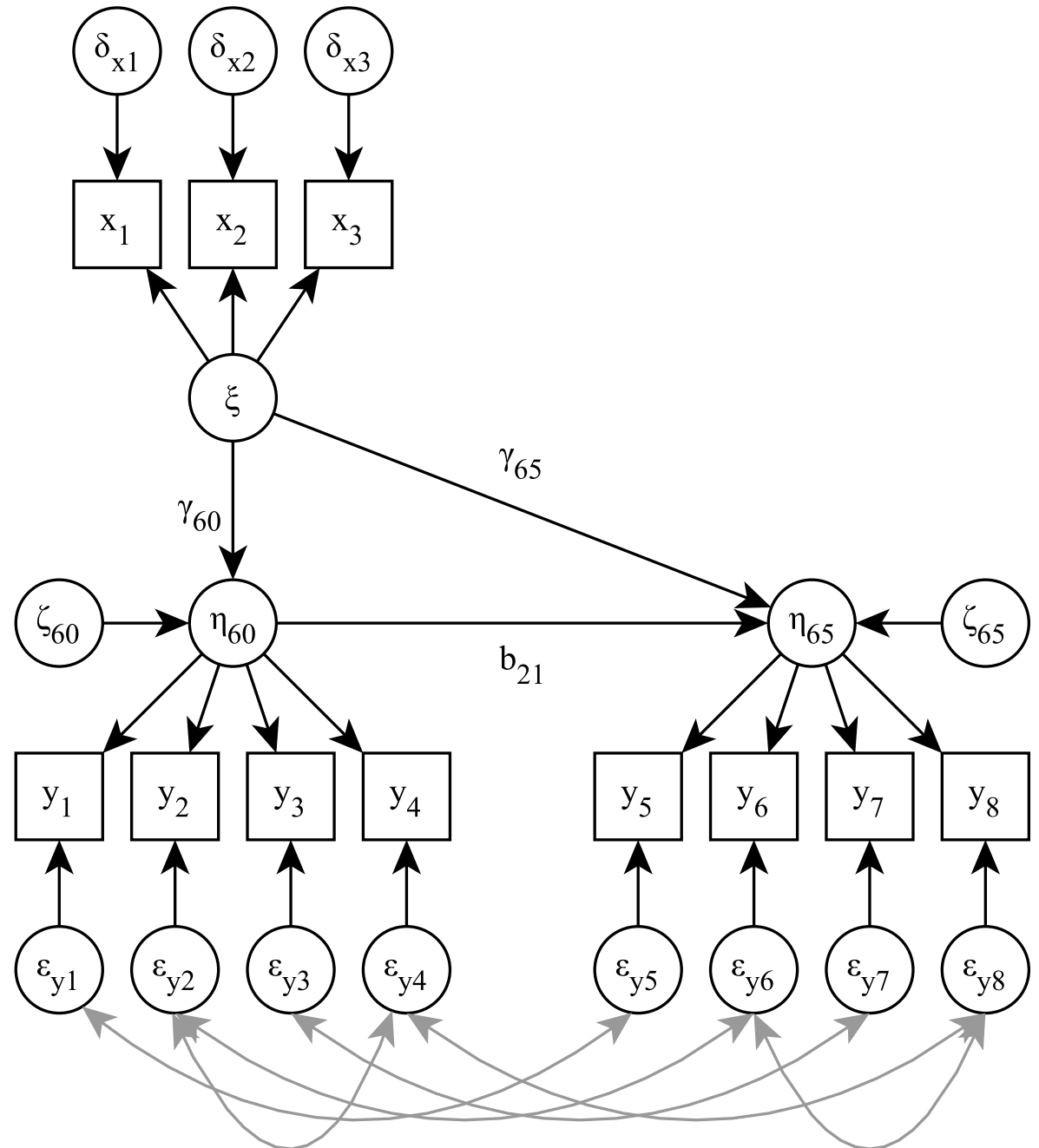
Results of interest do not differ across priors. If results differ, this provides valuable information.

Difficulties

- Models with many parameters
- Which priors to include

SEM example

What is the indirect effect of industrialization in 1960 on political democracy in 1965?



Prior sensitivity analysis: Which parameters?

- Focus on parameters of interest
- Latent variable variances are often sensitive

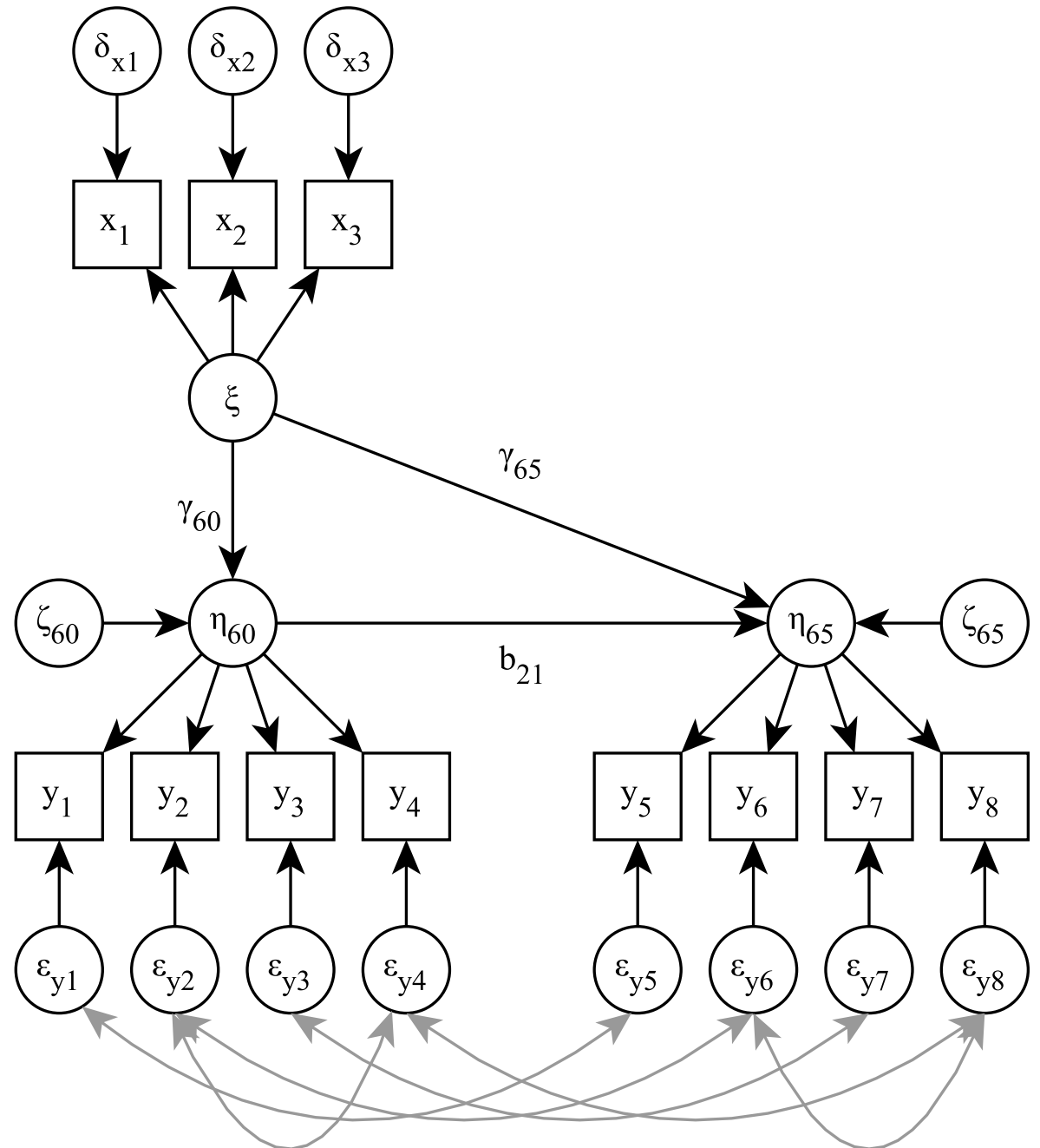
ϵ_{y7} δ_{x1} b_{21} ϵ_{y8}
 ϵ_{y4} ϵ_{y2} ξ ϵ_{y6}
 ϵ_{y5} η_{65} γ_{60} δ_{x3} η_{60}
 ζ_{60} ϵ_{y1} γ_{65} δ_{x2} η_{65} ϵ_{y3}

SEM example

Parameters of interest

- Indirect effect $\gamma_{60} b_{21}$
- Direct effect γ_{65}

Latent variable variances



Prior sensitivity analysis: Which priors?

- Distributional form depends on the parameter type
- Software can limit the possibilities
- When the original priors were informative: compare to default priors to see the influence of the informative priors and possibly to other levels of informativeness to be certain of your prior
- When the original priors were “non-informative” or default choices: compare to other default choices to ensure your priors are truly non-informative!

SEM example

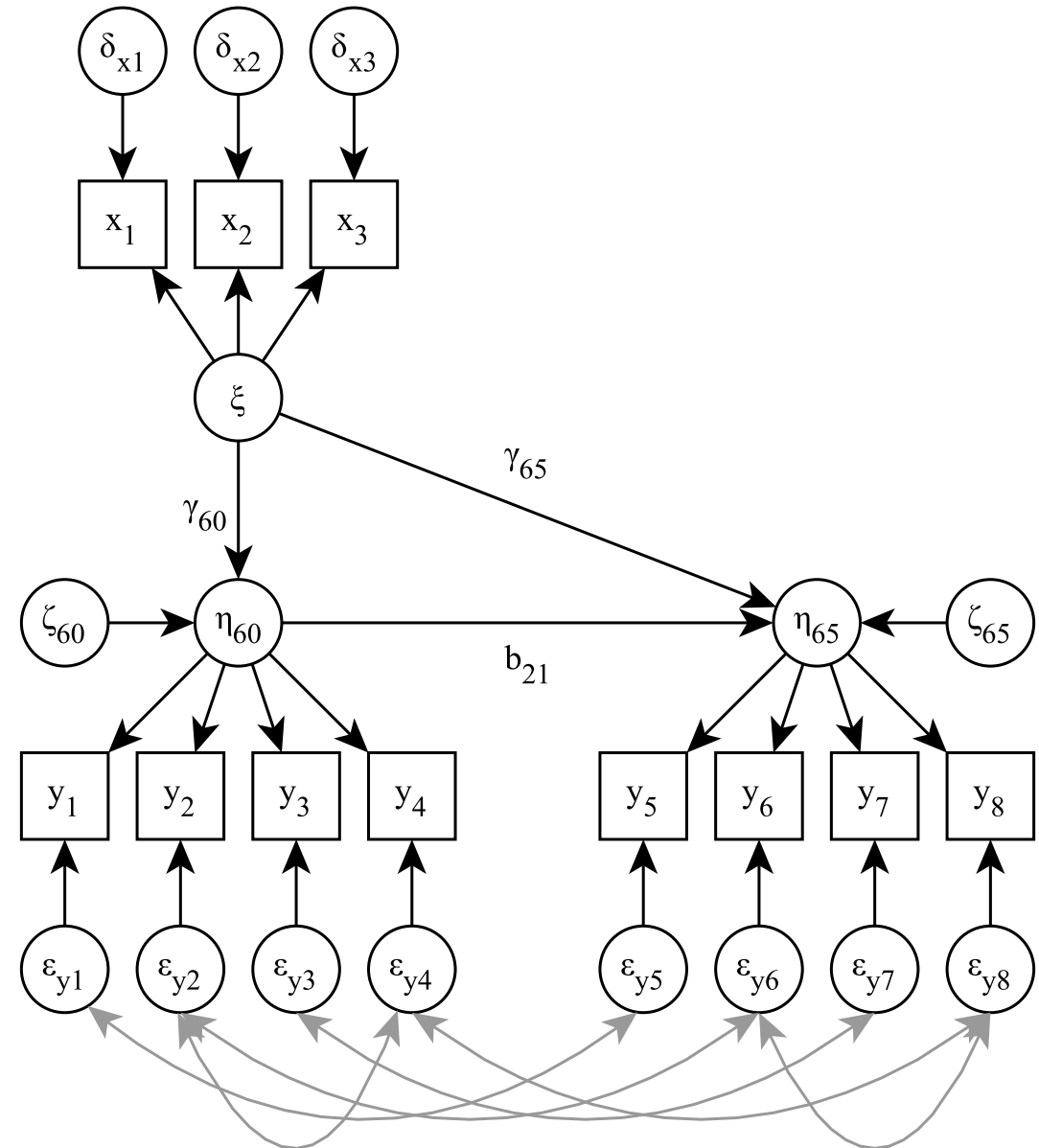
Original priors/baseline

Mplus defaults $N(0, 10^{10})$ & $\pi(\sigma^2) \propto 1$

Comparison

- $\pi(\sigma^2) \propto \sigma^{-1}$ ($IG(-0.5, 0)$)
- $\pi(\sigma^2) \propto \sigma^{-2}$ ($IG(0, 0)$)
- $\pi(\sigma^2) \propto IG(.1, .1)$
- $\pi(\sigma^2) \propto IG(.01, .01)$
- $\pi(\sigma^2) \propto IG(.001, .001)$
- Informative priors

Based on: van Erp, Mulder & Oberski (2018)



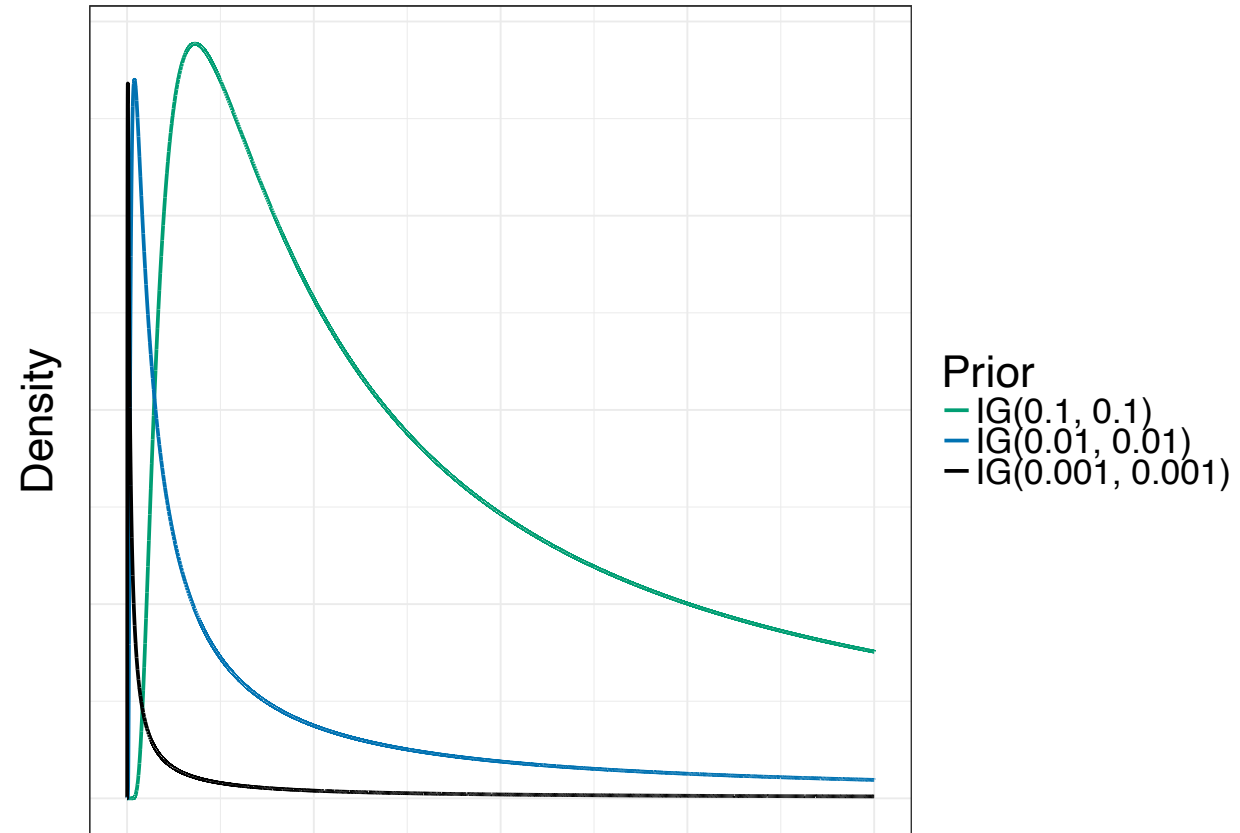
Some default priors for variances

Improper priors

1. Uniform on the variance:
 $\pi(\sigma^2) \propto 1 \rightarrow IG(-1, 0)$
2. Uniform on the SD:
 $\pi(\sigma^2) \propto \sigma^{-1} \rightarrow IG(-0.5, 0)$
3. Uniform on the log(var):
 $\pi(\sigma^2) \propto \sigma^{-2} \rightarrow IG(0, 0)$

Main issue: Can lead to improper posteriors

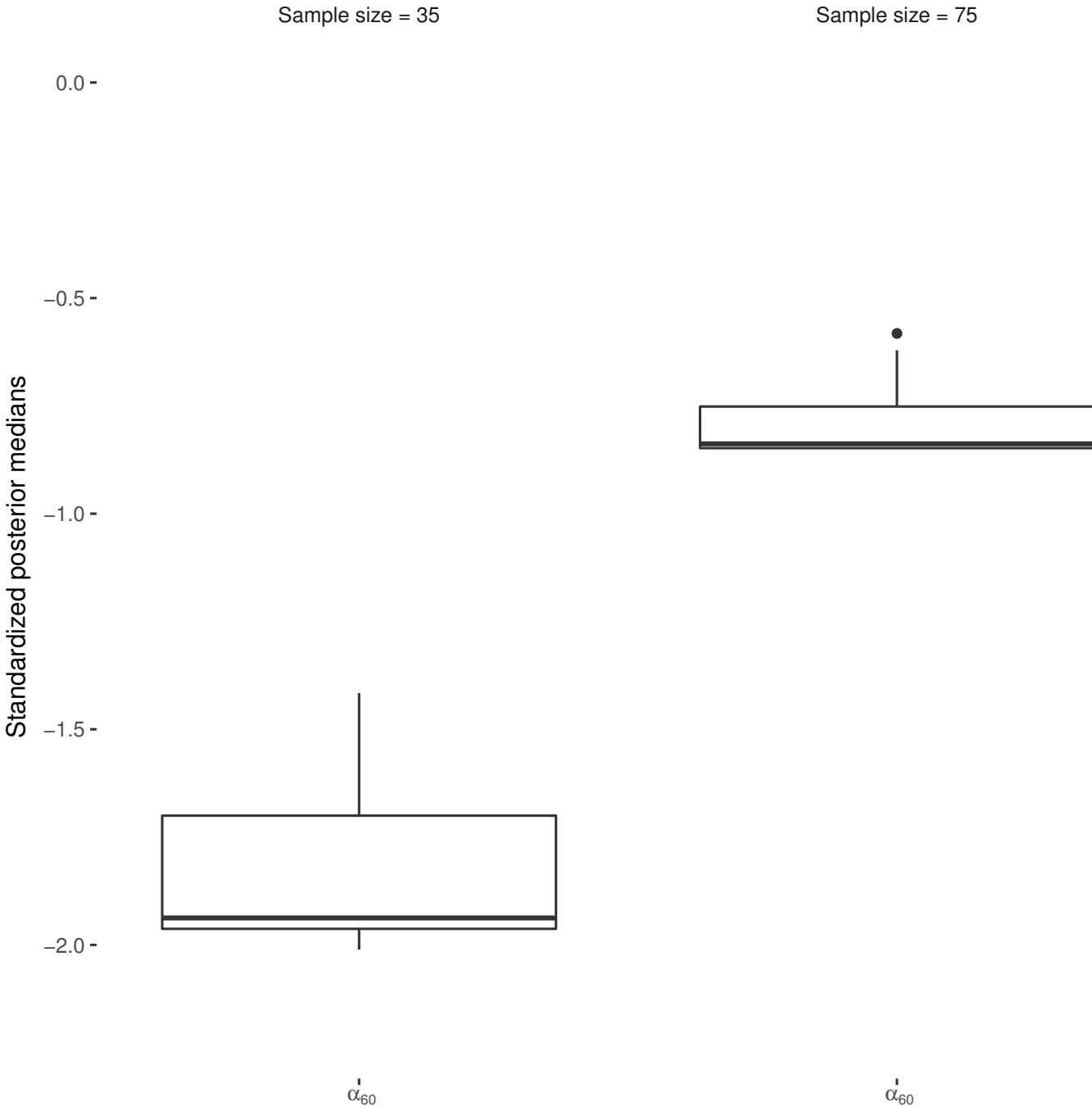
See also Gelman (2006) and van Erp & Browne (2021)



Prior sensitivity analysis: Practical considerations

- Depending on the number of analyses, convergence cannot be checked as extensively
- Depending on the number of parameters, change priors on groups of parameters simultaneously
- Consider beforehand what a meaningful change in a parameter would be
- **The goal is to ensure robust results, so be as critical as possible!**
Prior sensitivity is not necessarily bad, it's a source of information.
- Never change your prior afterwards to get the "best" results

Prior sensitivity analysis: Results



Prior sensitivity analysis: Results

Standardized and Unstandardized Point Estimates and 95% Confidence and Credible Intervals for the Direct Effect γ_{65} in the Prior Sensitivity Analysis

Prior	Standardized estimate	Unstandardized estimate	Lower bound 95% CI	Upper bound 95% CI	Width 95% CI
Sample size = 35					
Mplus default	.299	1.137	.132	2.193	2.061
$\pi(\sigma^2) \propto \sigma^{-1}$.284	1.052	.090	2.087	1.997
IG(.001, .001)	.270	.990	.074	2.041	1.967
IG(.01, .01)	.278	1.019	.059	2.029	1.970
IG(.1, .1)	.283	1.052	.088	2.053	1.965
Vague normal	.293	1.085	.114	2.086	1.972
EB1	.270	.975	.160	1.741	1.581
EB2	.274	.997	.137	1.812	1.675
Informative	.090	.225	-.427	.791	1.218

Prior sensitivity analysis: Results

1. Not sensitive
Robust
2. Default priors do not vary, but informative priors do
Prior knowledge has an influence -> is your prior an accurate representation of your beliefs?
3. Results vary across all priors, incl. defaults
Small sample -> collect more data or report the range of results

Make sure you are transparent and report all steps and results from your sensitivity analysis!



Part 2: Shrinkage priors

Example: Predicting the number of murders

- Suppose we wish to predict the number of murders in US communities.
- We have 125 predictors.
- We need at least 125 communities to fit the model.
- Even with 126 we would be likely overfitting.

General

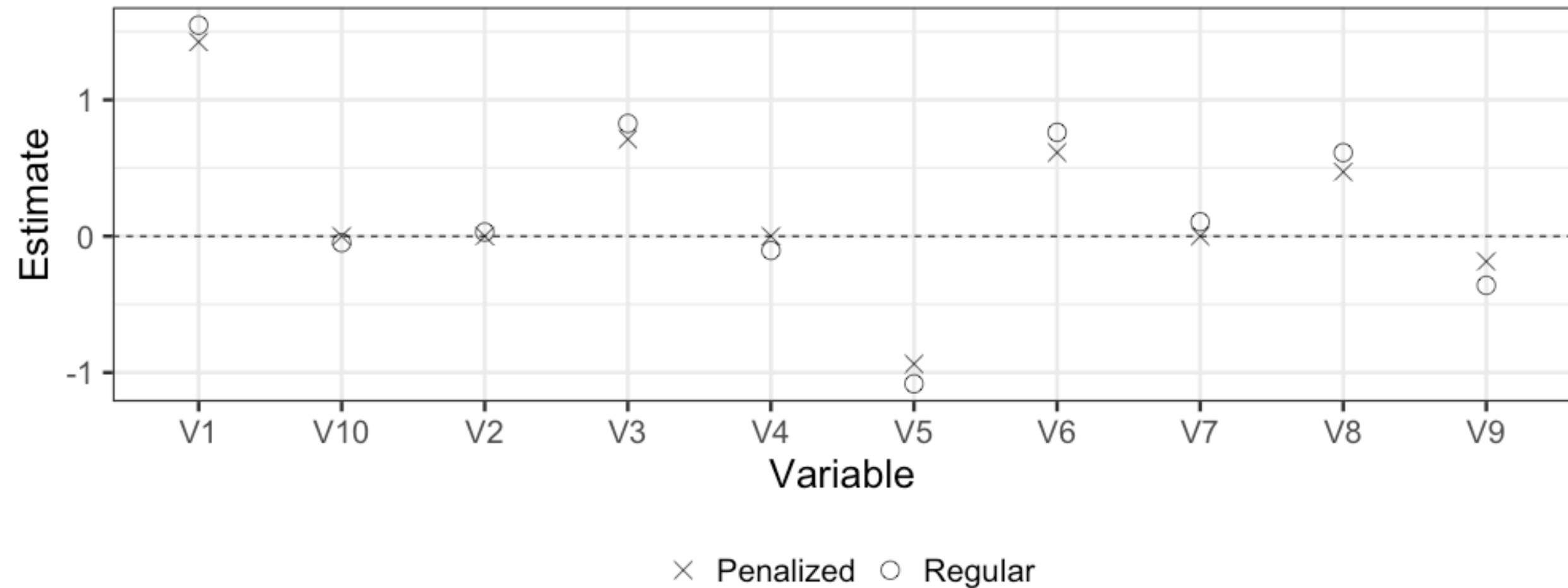
We want a big enough n to p ratio.
What if this is not the case?



Regularized/penalized regression

- Add a penalty term to OLS, e.g., lasso, ridge or elastic net
- This will shrink small coefficients to zero
- Some penalties also perform variable selection
- Bias is introduced to avoid overfitting.

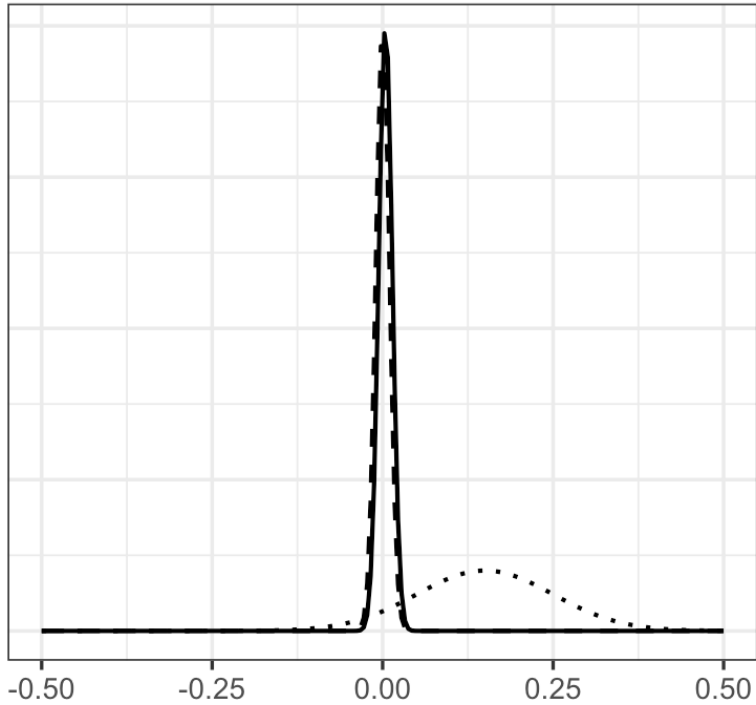
Illustration: lasso penalty



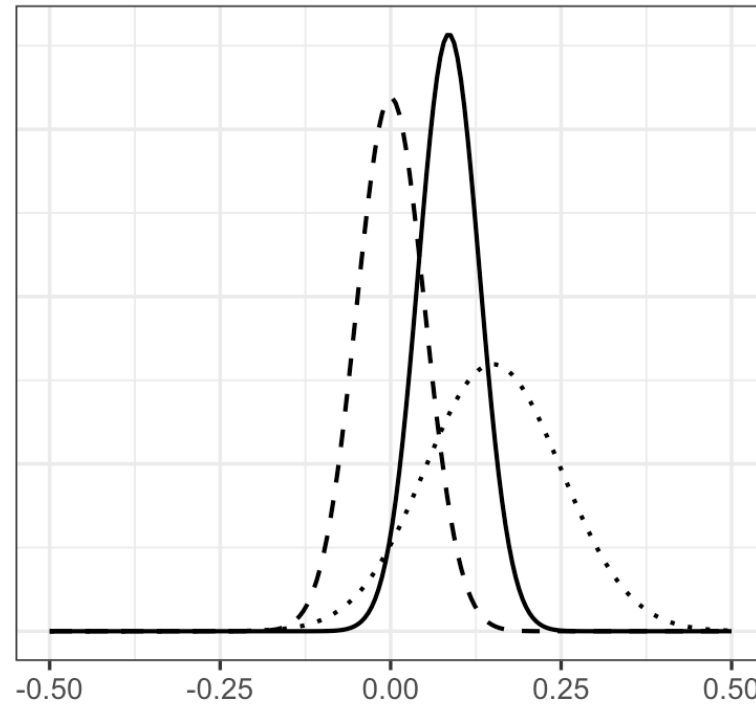
Bayesian regularization

Instead of using a penalty function, we use the **prior distribution!**

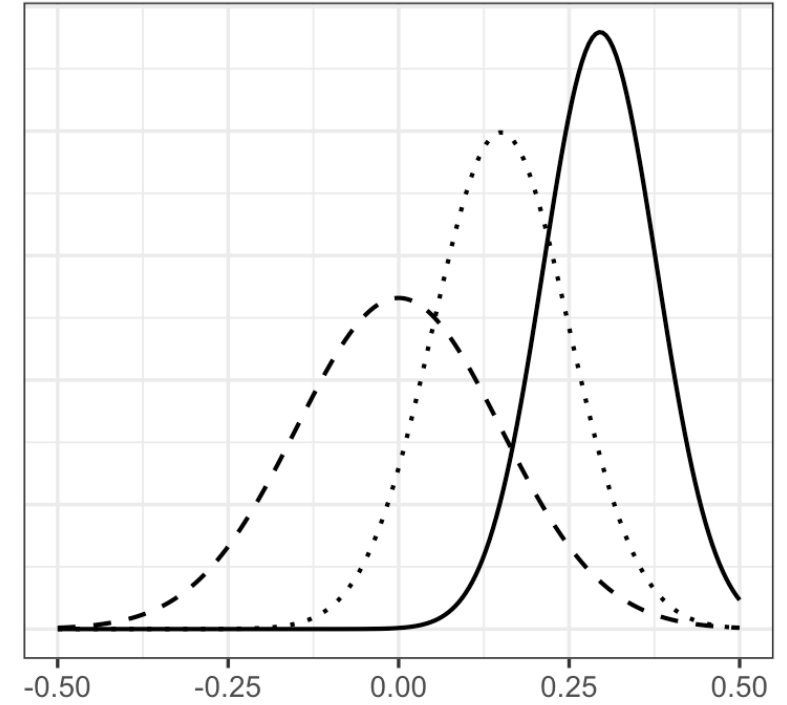
Prior standard deviation 0.01



Prior standard deviation 0.05



Prior standard deviation 0.15



Bayesian regularization

- Instead of using a penalty, we use the prior
- Specify the prior such that small effects are pulled to zero
- Ideally, substantial effects remain large
- Many different shrinkage priors try this
- Some shrinkage priors correspond to classical penalty functions

Advantages Bayesian regularization

- Regularization comes naturally in the Bayesian framework

We need to specify a prior anyway

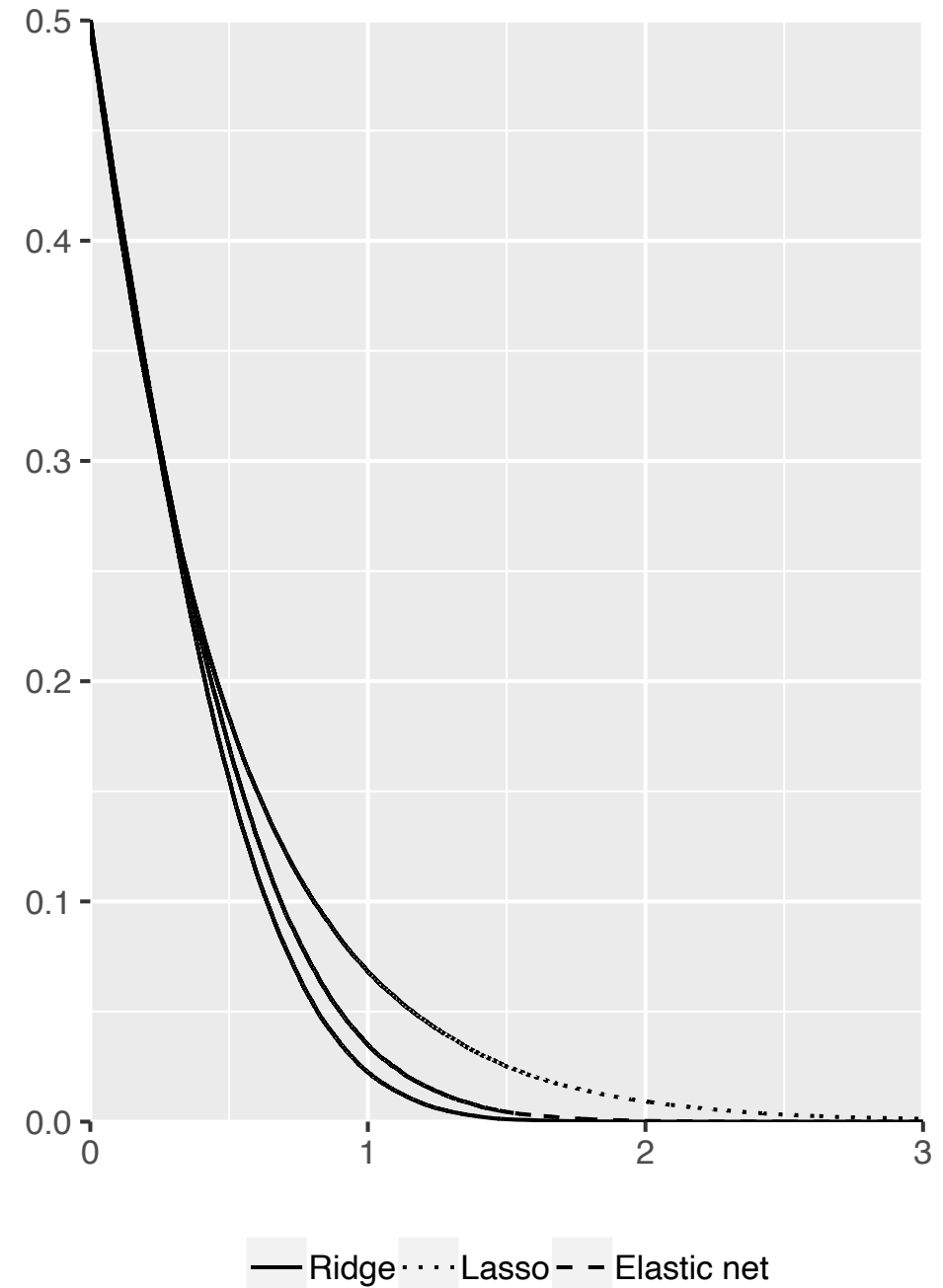
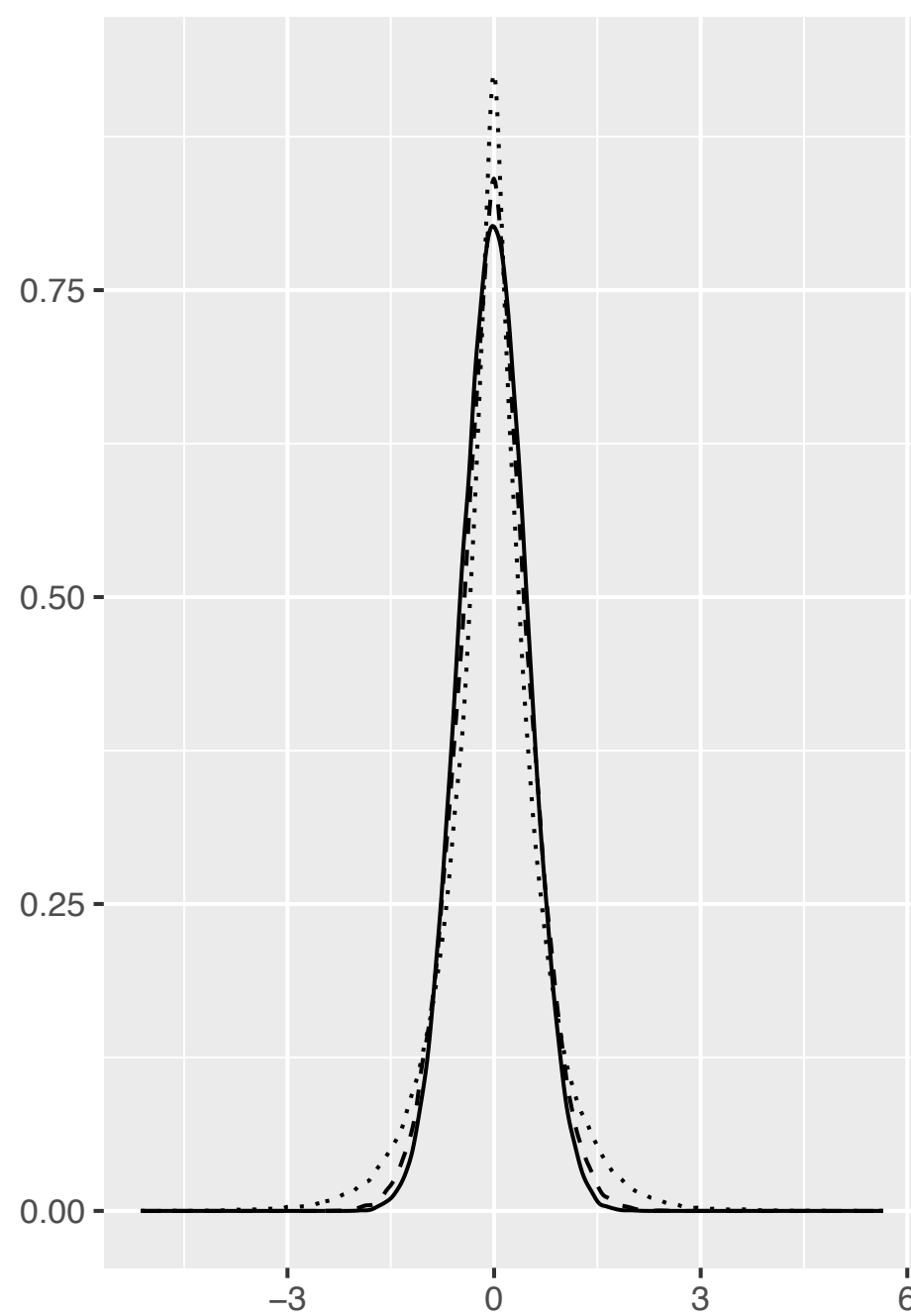
- Simultaneous estimation penalty parameter or the amount of shrinkage

Full Bayes approach

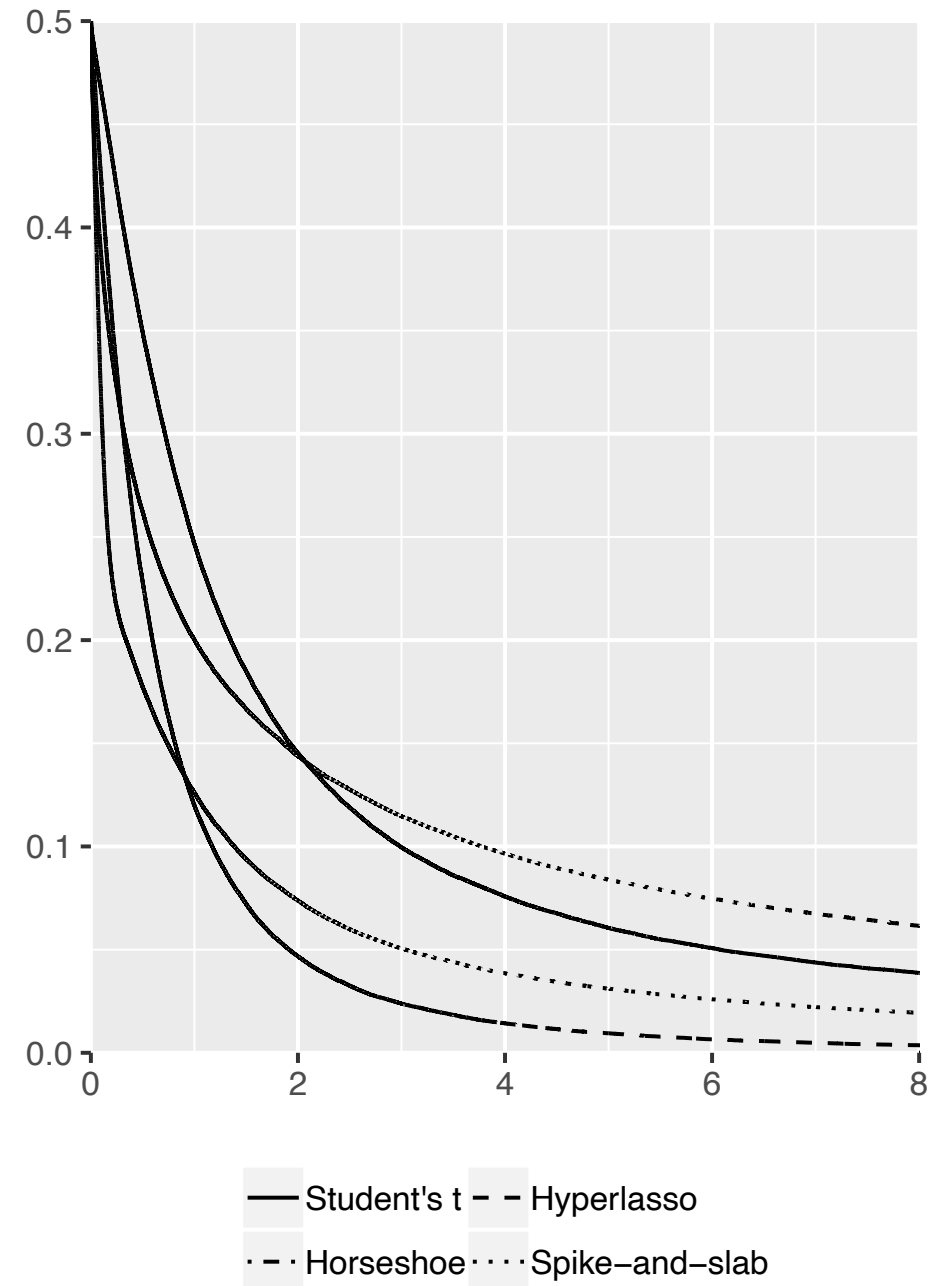
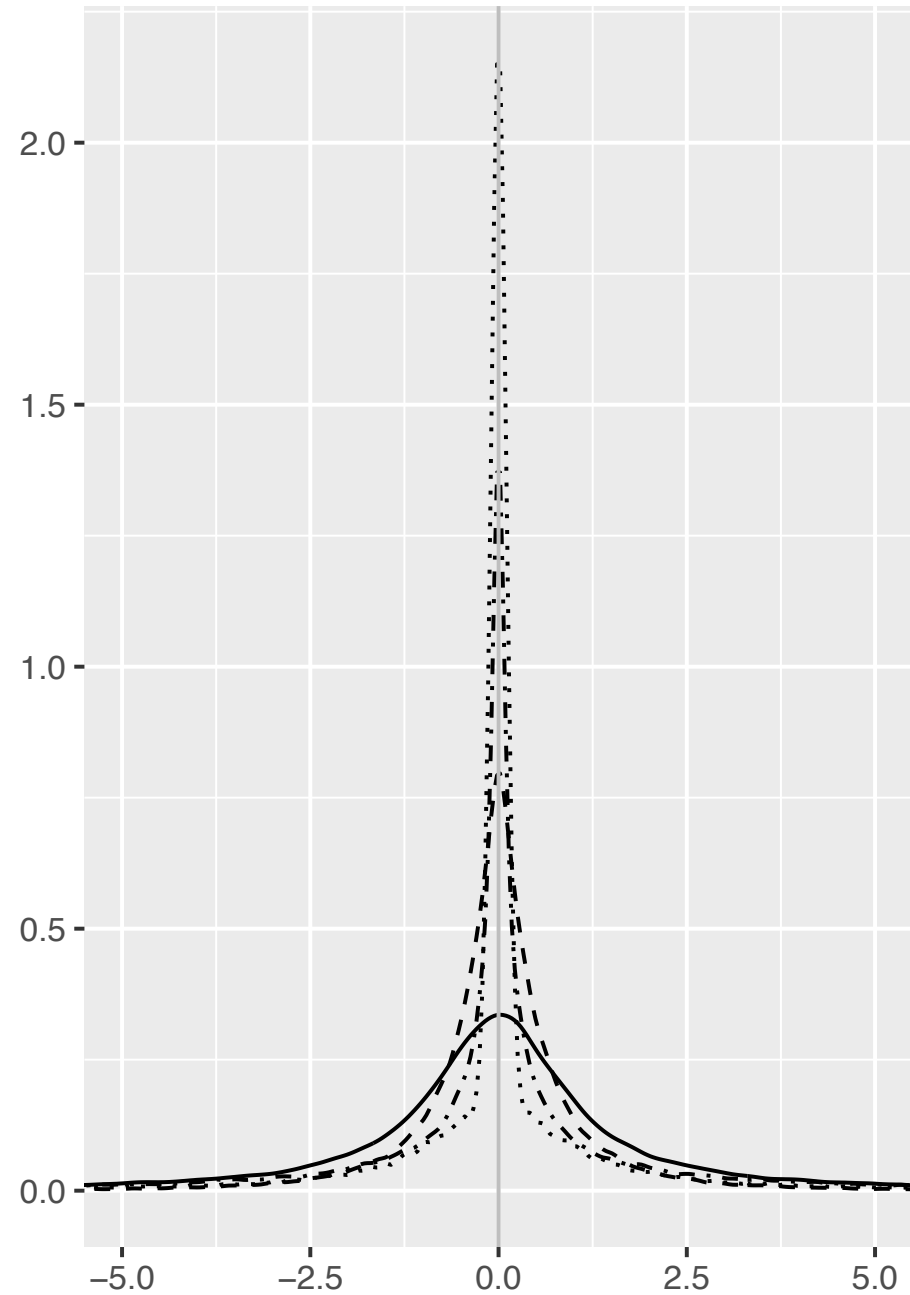
- Flexibility in terms of shrinkage priors

Including counterparts classical penalties

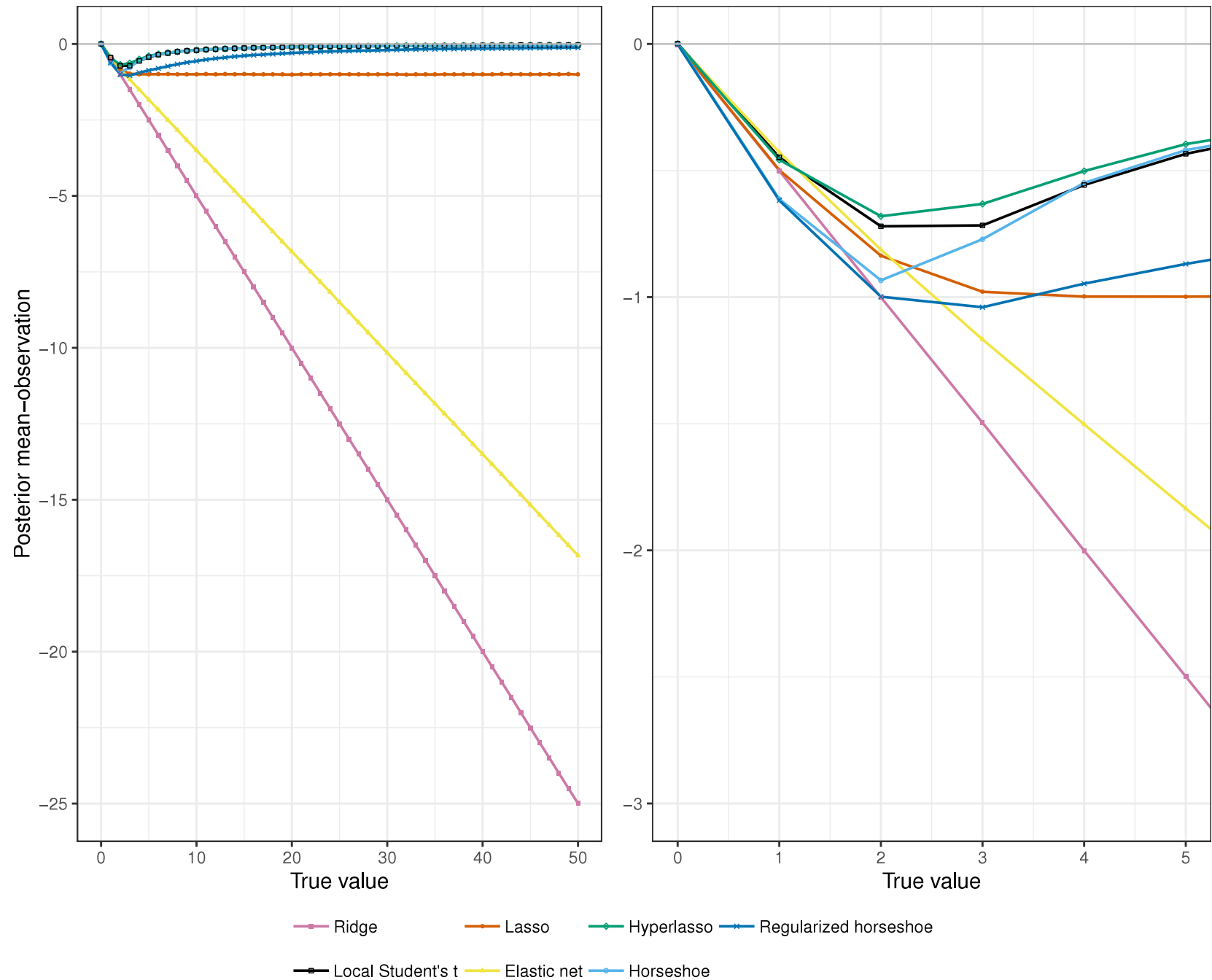
Many different shrinkage priors exist



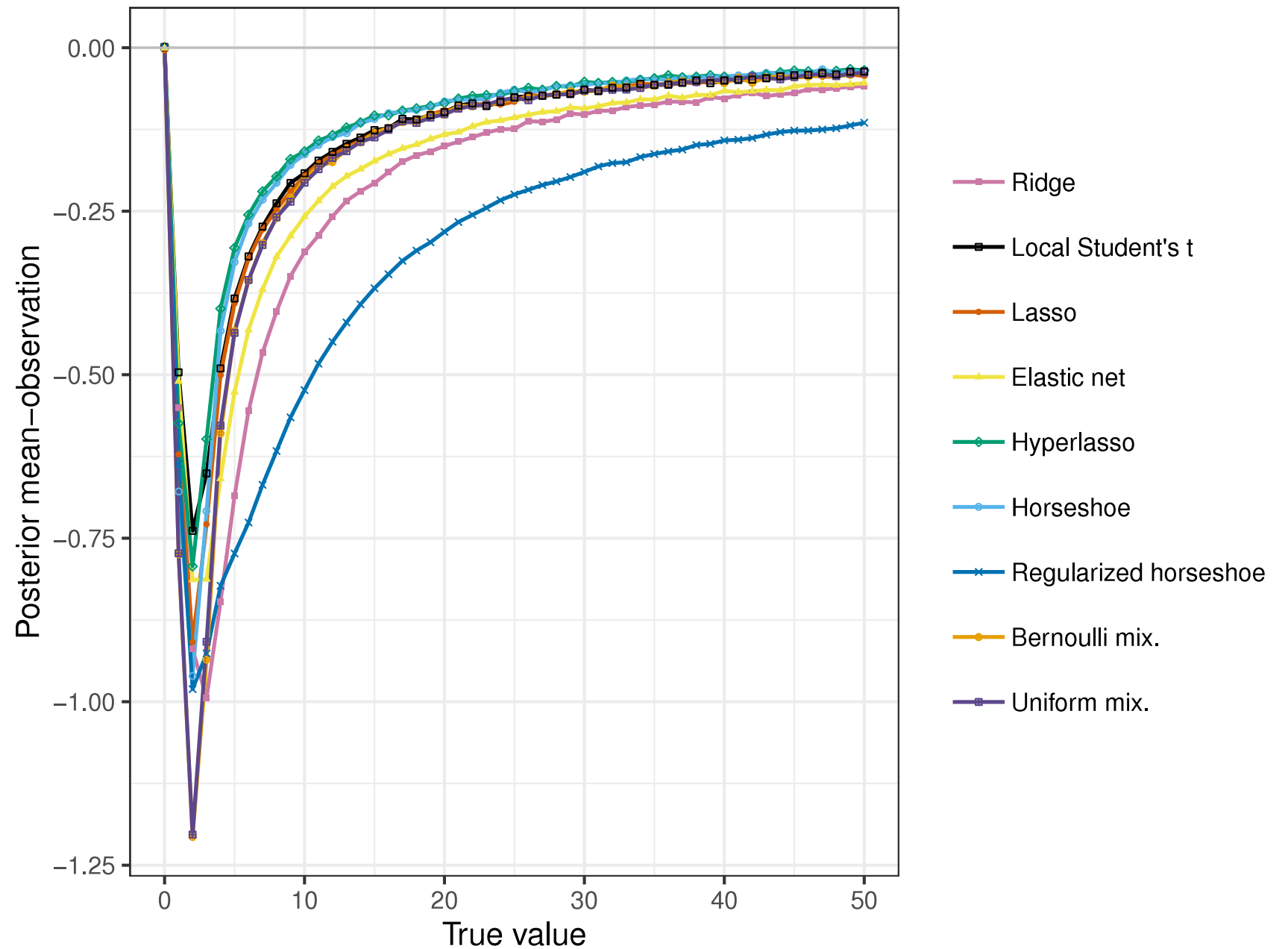
Many different shrinkage priors exist



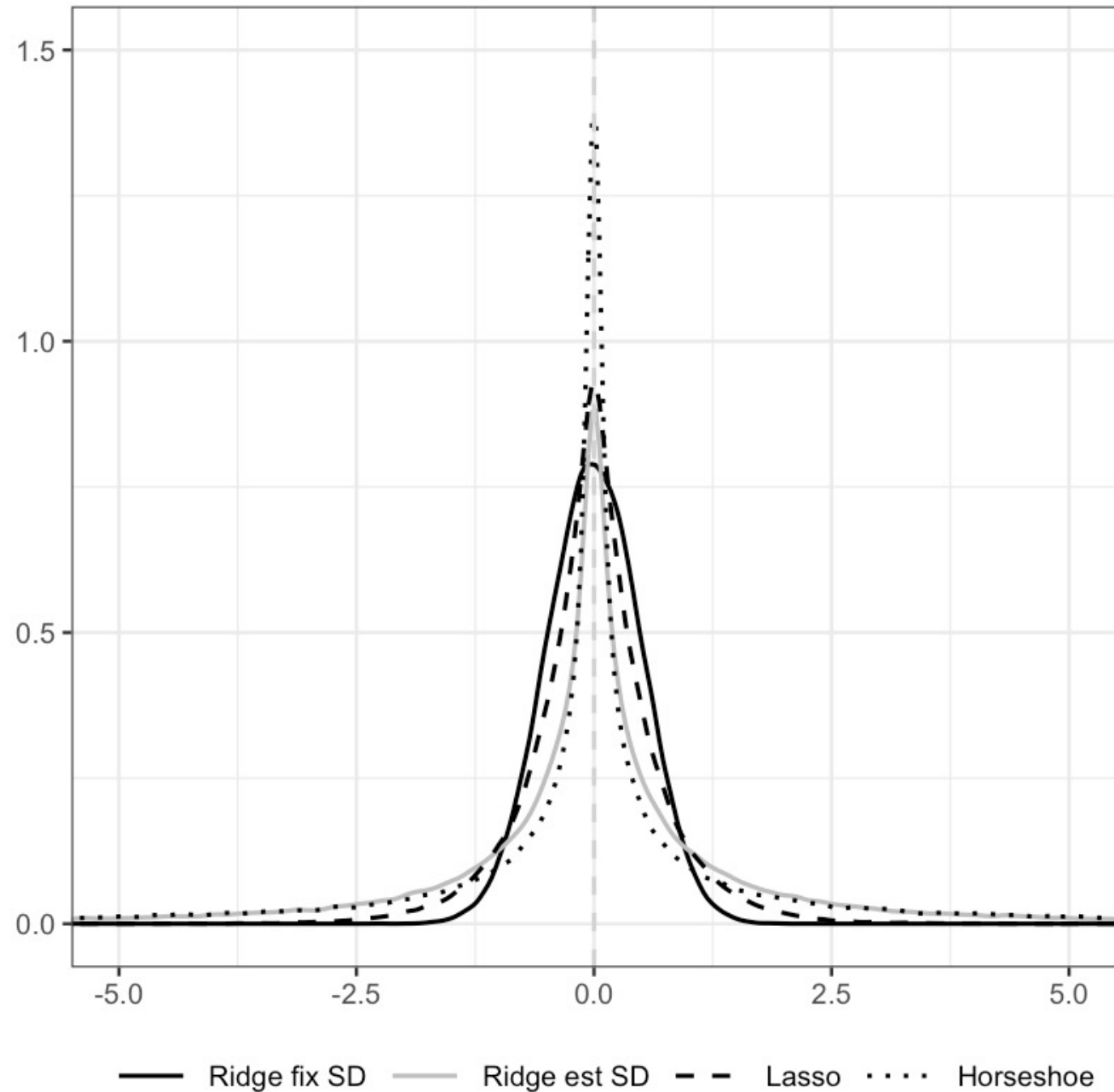
Leading to different shrinkage behaviors



Leading to
different
shrinkage
behaviors



Penalty parameter can be fixed or estimated



Penalty parameter?

Which shrinkage prior?

Practical considerations

Variable selection?

Determination penalty parameter

- Cross-validation (classical framework)
- Fixed value
- Empirical Bayes
- Full-Bayes

Software dependent, but full Bayes is generally most robust.

Choice shrinkage prior

Restricted by software

brms is quite flexible, including ridge, lasso, regularized horseshoe

Simpler priors easier to understand, complex priors might perform better

Generally: most priors perform similarly when $p < n$. For more complex models, more advanced priors might be more suitable

Visualizations and prior sensitivity analyses can provide insight

How to select parameters

Classical lasso automatically sets parameters to zero

Bayesian point estimates are never exactly zero

Potential variable selection criteria:

1. Cut-off value (e.g., 0.1)
2. Credible interval
3. Projection predictive variable selection

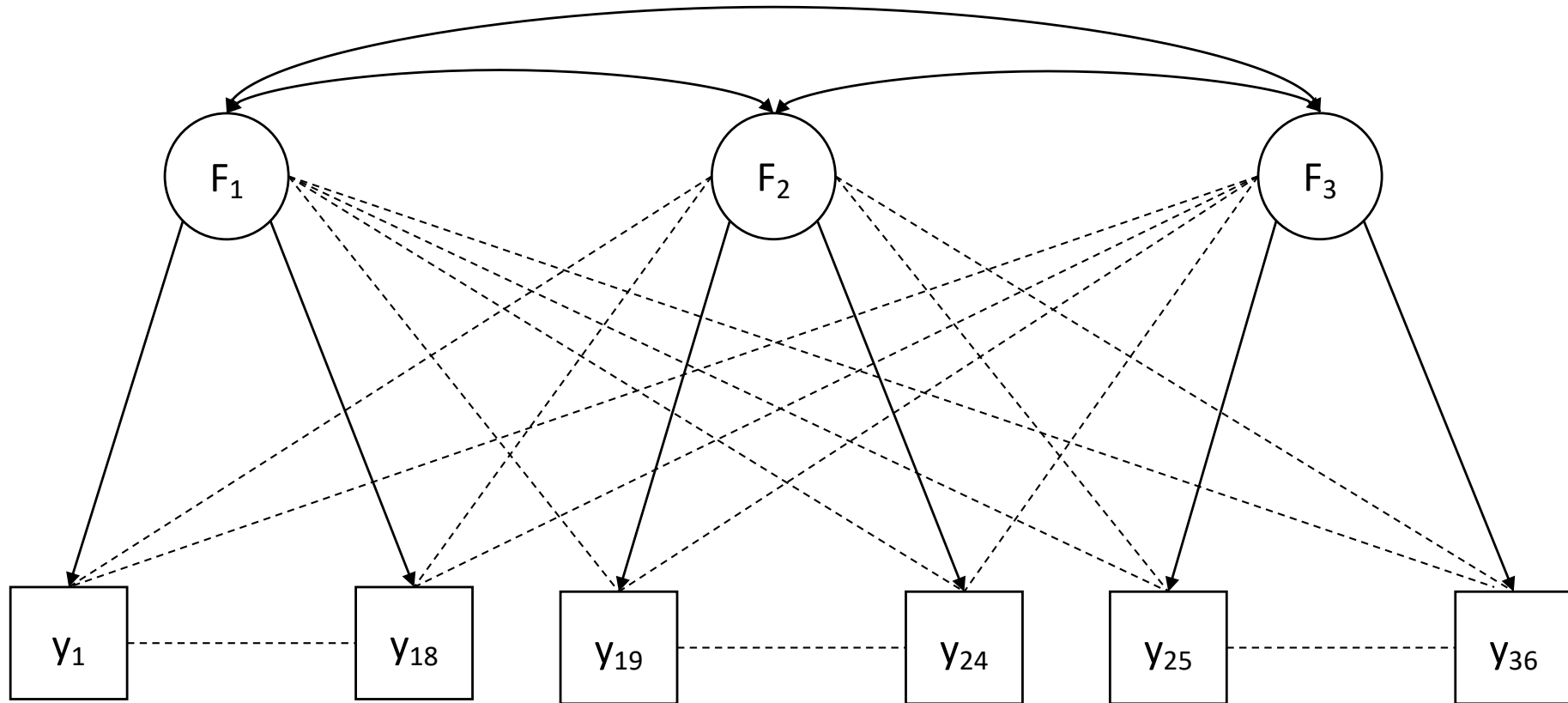
Beyond regression models

(Bayesian) regularization can be used in any model where we can assume a priori that some parameters equal zero:

E.g., to select moderators in meta-analysis (van Lissa, van Erp, & Clapper, 2023)

Or in structural equation modeling (van Erp, 2023)

Bayesian regularized SEM



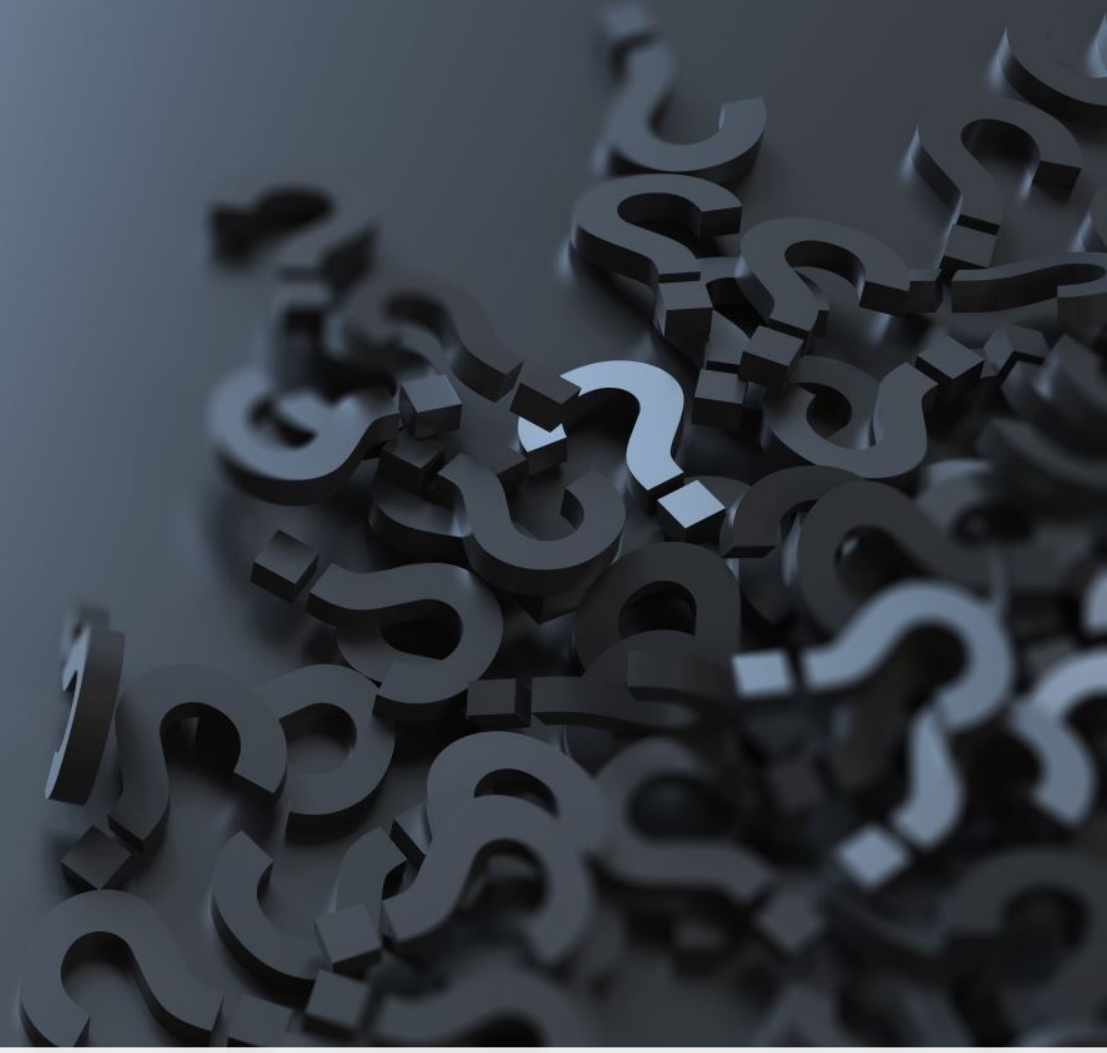
Recap

Part 1: Prior sensitivity analysis

- Recap: What is a prior?
- When is a prior influential?
- How to perform a prior sensitivity analysis

Part 2: Shrinkage priors

- Basic idea behind penalization
- Different shrinkage priors = different behaviors
- Practical considerations
- Advanced applications



Questions?