Economic predictions with big data: The illusion of sparsity

Domenico Giannone
New York Fed

Michele Lenza
European Central Bank

Giorgio Primiceri
Northwestern University

10th ECB Workshop on Forecasting Techniques

Frankfurt, June 19, 2018

Predictive modeling with big data

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t\,,\;\; \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Predictive modeling with big data

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t\,, \;\; \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Big data: large k

Predictive modeling with big data

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t\ ,\ \ \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Big data: large k

- Standard inference (ML or flat prior) is a bad idea
  - Proliferation of parameters
  - High estimation uncertainty
  - Overfitting and imprecise out-of-sample forecasting / poor external validity

Predictive modeling with big data

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t \,, \;\; \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Big data: large k

- Standard inference (ML or flat prior) is a bad idea
  - Proliferation of parameters
  - High estimation uncertainty
  - Overfitting and imprecise out-of-sample forecasting / poor external validity

➥ Methods to address curse of dimensionality (Ng, 2013, CHL, 2017)
  - **Sparse** modeling e.g. hand picking, Lasso regression
  - **Dense** modeling e.g. Ridge regression, Factor models

This paper: Sparse or dense modeling?

This paper: Sparse or dense modeling?

- Answer is an empirical matter
  - Study a variety of predictive problems in macro, micro and finance

This paper: Sparse or dense modeling?

- Answer is an empirical matter
  - Study a variety of predictive problems in macro, micro and finance

- Popular techniques not suitable to answer the question
  - Sparsity/density often assumed
  - A small set of predictors might be selected simply to reduce estimation error, even if the model is not sparse

This paper: Sparse or dense modeling?

- Answer is an empirical matter
  - Study a variety of predictive problems in macro, micro and finance

- Popular techniques not suitable to answer the question
  - Sparsity/density often assumed
  - A small set of predictors might be selected simply to reduce estimation error, even if the model is not sparse

- Our predictive model
  - **sparsity**, without assuming it
  - **shrinkage**, to give a chance to large models
  - Bayesian inference on sparsity and shrinkage

Main results

1. No clear pattern of sparsity
   - Posterior not concentrated on a single sparse model, but on a wide set

Main results

1. No clear pattern of sparsity
   - Posterior not concentrated on a single sparse model, but on a wide set

2. More sparsity emerges only if very tight prior favoring small models

Main results

1. No clear pattern of sparsity
   - Posterior not concentrated on a single sparse model, but on a wide set

2. More sparsity emerges only if very tight prior favoring small models

The illusion of sparsity

Outline

- The predictive model
- Applications to macro, micro and finance
- Sparse or dense modeling?
  - Exploring the posterior

Outline

- The predictive model
- Applications to macro, micro and finance
- Sparse or dense modeling?
  - Exploring the posterior

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t \,, \;\; \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t \,, \;\; \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} & \\ 0 & \text{with probability} \quad 1 - q \end{cases}$$

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t \,, \;\; \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} \mathcal{N}(0, \sigma^2 \gamma^2) & \text{with probability} \quad q \\ 0 & \text{with probability} \quad 1 - q \end{cases}$$

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t\ ,\ \ \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} \mathcal{N}(0, \sigma^2\gamma^2) & \text{with probability} \quad \color{red}{q} \\ 0 & \text{with probability} \quad 1 - q \end{cases}$$

Probability of inclusion, controls **size**

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t \,, \;\; \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} \mathcal{N}(0, \sigma^2 \gamma^2) & \text{with probability} \quad q \\ 0 & \text{with probability} \quad 1 - q \end{cases}$$

Variance of prior, controls **shrinkage**

Probability of inclusion, controls **size**

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t \,,\;\; \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} \mathcal{N}(0, \sigma^2 \gamma^2) & \text{with probability} \quad q \\ 0 & \text{with probability} \quad 1 - q \end{cases}$$

- "Spike-and-slab" prior
  - Mitchel and Beauchamp (1988)
  - Vast literature on Bayesian Model Averaging and Variable Selection

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t \ , \ \ \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} \mathcal{N}(0, \sigma^2 \gamma^2) & \text{with probability} \quad q \\ 0 & \text{with probability} \quad 1-q \end{cases}$$

- "Spike-and-slab" prior
  - Mitchel and Beauchamp (1988)
  - Vast literature on Bayesian Model Averaging and Variable Selection
  - This paper: inference on $\boldsymbol{q}$ and $\boldsymbol{\gamma^2}$

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t \, , \;\; \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} \mathcal{N}(0, \sigma^2\gamma^2) & \text{with probability} \quad q \\ 0 & \text{with probability} \quad 1-q \end{cases}$$

- Hyperpriors

$$q \sim \mathcal{B}(a, b)$$

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t \,, \;\; \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} \mathcal{N}(0, \sigma^2\gamma^2) & \text{with probability} \quad q \\ 0 & \text{with probability} \quad 1-q \end{cases}$$

- Hyperpriors

$$q \sim \mathcal{B}(a, b), \qquad \frac{q\ k\ var(x)\ \gamma^2}{q\ k\ var(x)\ \gamma^2 + 1} \equiv R^2$$

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t\ ,\ \ \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} \mathcal{N}(0, \sigma^2 \gamma^2) & \text{with probability} \quad q \\ 0 & \text{with probability} \quad 1-q \end{cases}$$

- Hyperpriors

$\propto E[var(y_t) | q, \gamma^2, \sigma^2]$ $\propto E[var(x_t'\beta) | q, \gamma^2, \sigma^2]$

$$q \sim \mathcal{B}(a, b), \qquad \frac{q\ k\ var(x)\ \gamma^2}{q\ k\ var(x)\ \gamma^2 + 1} \equiv R^2$$

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t \,, \;\; \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} \mathcal{N}(0, \sigma^2 \gamma^2) & \text{with probability} \quad q \\ 0 & \text{with probability} \quad 1 - q \end{cases}$$

- Hyperpriors

$$q \sim \mathcal{B}(a, b), \qquad \frac{q\ k\ var(x)\ \gamma^2}{q\ k\ var(x)\ \gamma^2 + 1} \equiv R^2 \sim \mathcal{B}(A, B)$$

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t\ ,\ \ \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

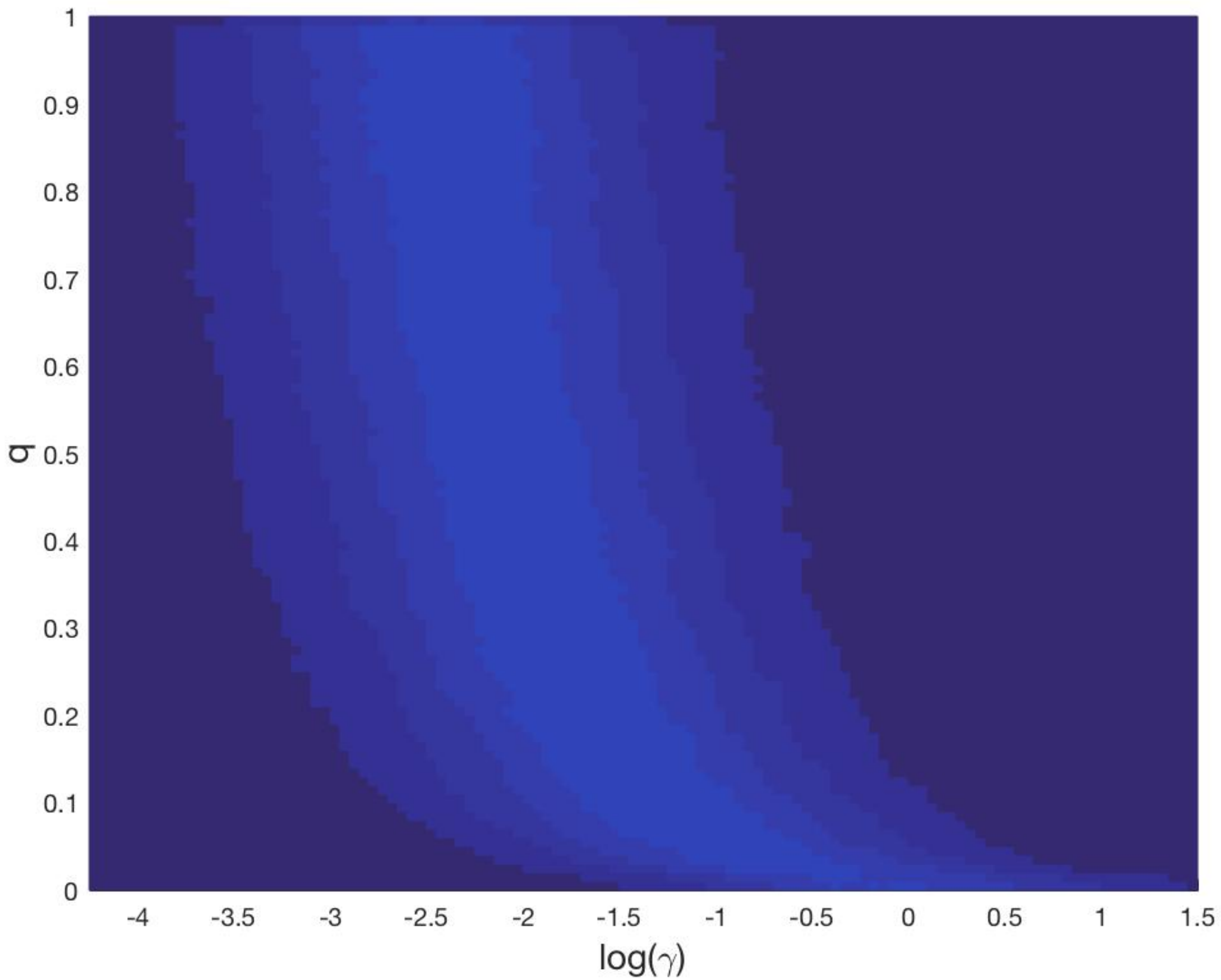
- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} \mathcal{N}(0, \sigma^2 \gamma^2) & \text{with probability} \quad q \\ 0 & \text{with probability} \quad 1-q \end{cases}$$

- Hyperpriors

$$q \sim \mathcal{B}(1, 1), \qquad \frac{q\ k\ var(x)\ \gamma^2}{q\ k\ var(x)\ \gamma^2 + 1} \equiv R^2 \sim \mathcal{B}(1, 1)$$

The implied joint prior on q and γ^2



Outline

- The predictive model
- Applications to macro, micro and finance
- Sparse or dense modeling?
  - Exploring the posterior

Economic applications

- Macro
  - Forecasting industrial production with many macro predictors
  - The determinants of economic growth in a cross-section of countries

- Finance
  - Prediction of the US aggregate equity premium over time
  - Explaining the cross-section of equity returns across firms

- Micro
  - Understanding the decline in crime rates in US states during the 1990s
  - The determinants of government takings of private properties in US judicial circuits

- Some references:
  - Stock-Watson (2002a and b), Barro-Lee (1994), Sala-i-Martin et al. (2004), Welch-Goyal (2008), Freyberger et al. (2017), Donohue-Levitt (2001), Chen-Yeh (2012), Belloni et al. (2011, 2012, 2014).

Economic applications

	Y	X	Sample
Macro 1	Growth rate of US **Industrial Prod.**	**130** lagged macro and financial indicators	**659** **time-series** obs. Feb. 60-Dec. 14
Macro 2	**Countries average growth** 1960-1985	**60** country charact' socio-econ, inst.	**90** **cross-section** obs.
Finance 1	US **equity premium**	**16** lagged macro and financial indicators	**58** **time-series** obs. 1948-2015
Finance 2	**Stock returns** of US firms	**144** dummies lagged characts'	≈**1400k** **panel** obs. Jul. 63–Dec. 15, ≈2k firms
Micro 1	**Crime rate** in US states	**285** state characts' socio-econ, inst., law	**476** **panel** obs. Jan. 86–Dec. 97, 48 states
Micro 2	**Eminent domain** judicial decisions	**138** judges' characts' socio-polit., profess	**312** **panel** obs. 1975-2008, circuits

Outline

- The predictive model
- Applications to macro, micro and finance
- Sparse or dense modeling?
  - Exploring the posterior

Exploring the posterior

1. No clear pattern of sparsity
    - Posterior not concentrated on a single sparse model, but on a wide set

2. More sparsity emerges only if very tight prior favoring small models

Exploring the posterior

0. Inclusion probability and shrinkage are complements, but imperfect

1. No clear pattern of sparsity
   - Posterior not concentrated on a single sparse model, but on a wide set

2. More sparsity emerges only if very tight prior favoring small models

Exploring the posterior

0. Inclusion probability and shrinkage are complements, but imperfect

1. No clear pattern of sparsity
   - Posterior not concentrated on a single sparse model, but on a wide set

2. More sparsity emerges only if very tight prior favoring small models

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t\ ,\ \ \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

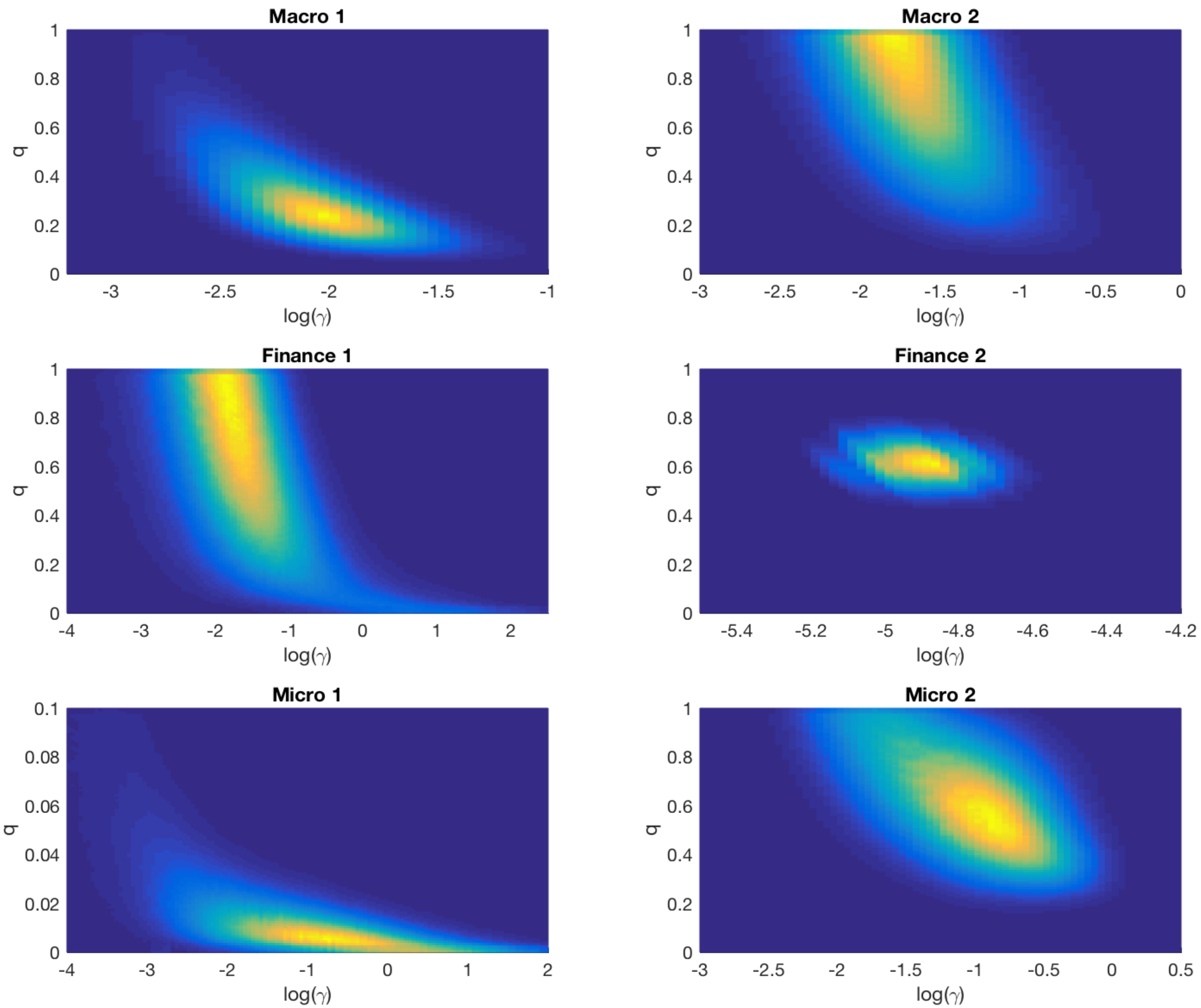
- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} \mathcal{N}(0, \sigma^2 \gamma^2) & \text{with probability} \quad q \\ 0 & \text{with probability} \quad 1 - q \end{cases}$$

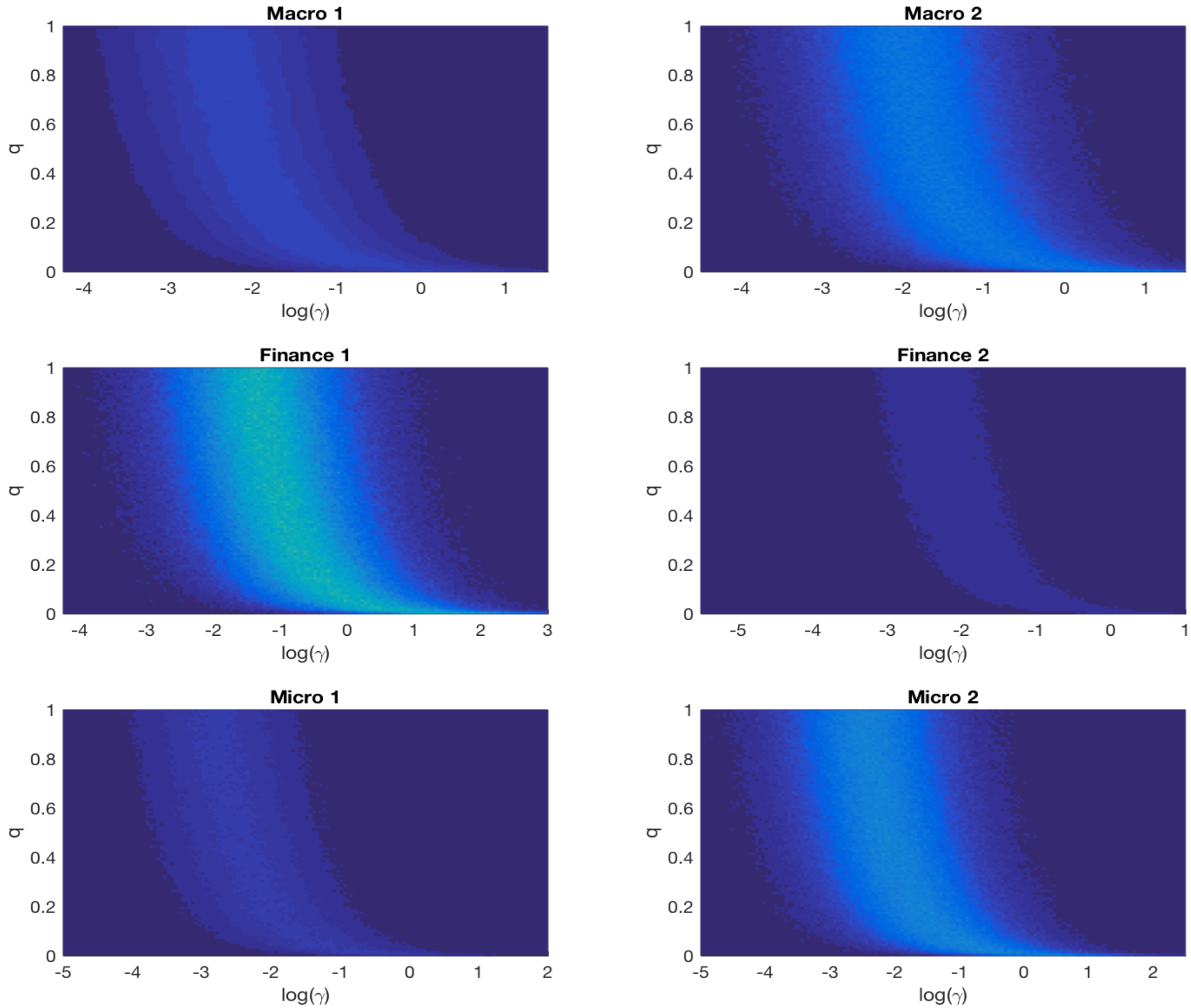
Variance of prior, controls **shrinkage**

Probability of inclusion, controls **size**

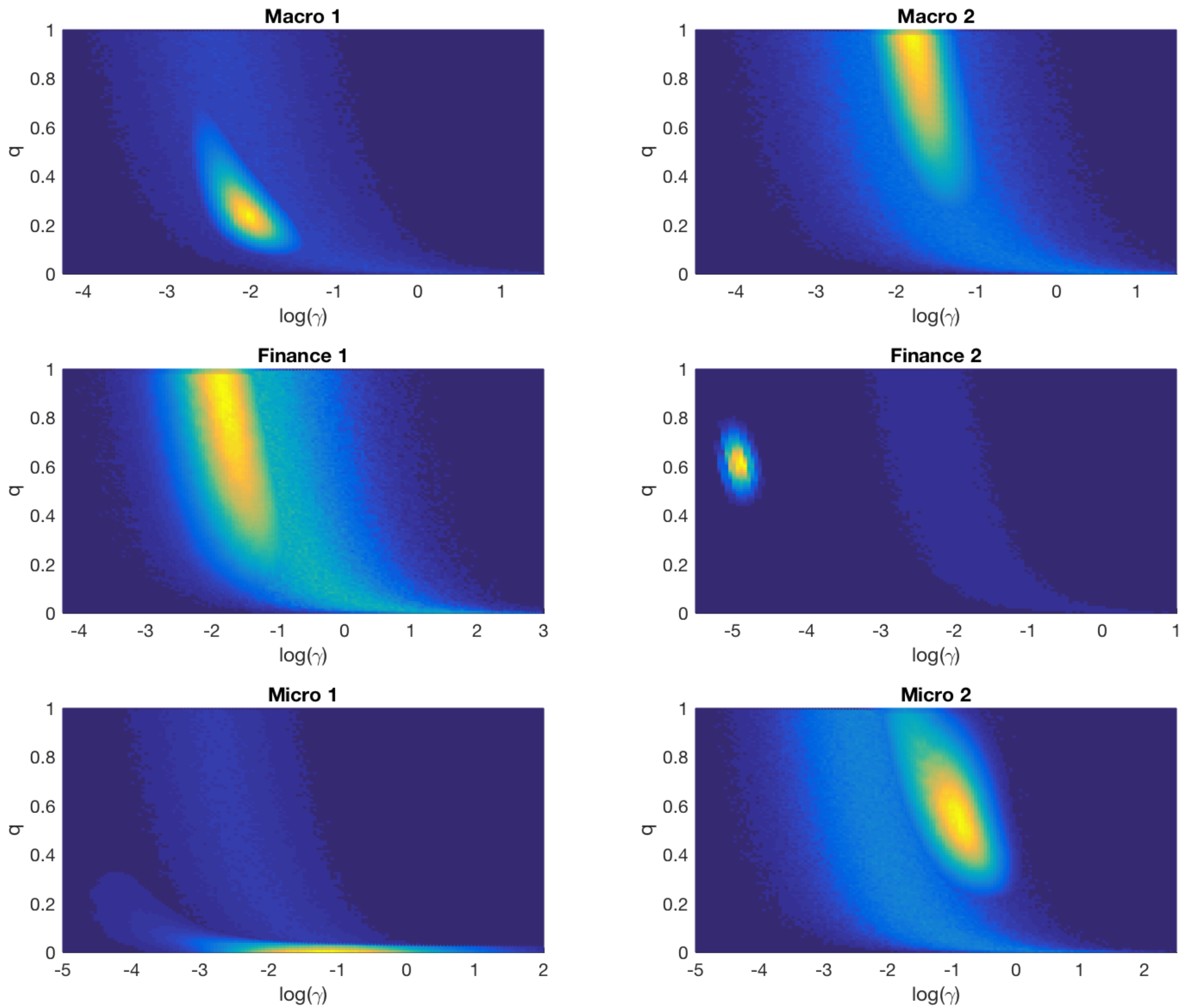
Probability of inclusion and prior variance (q and γ^2)



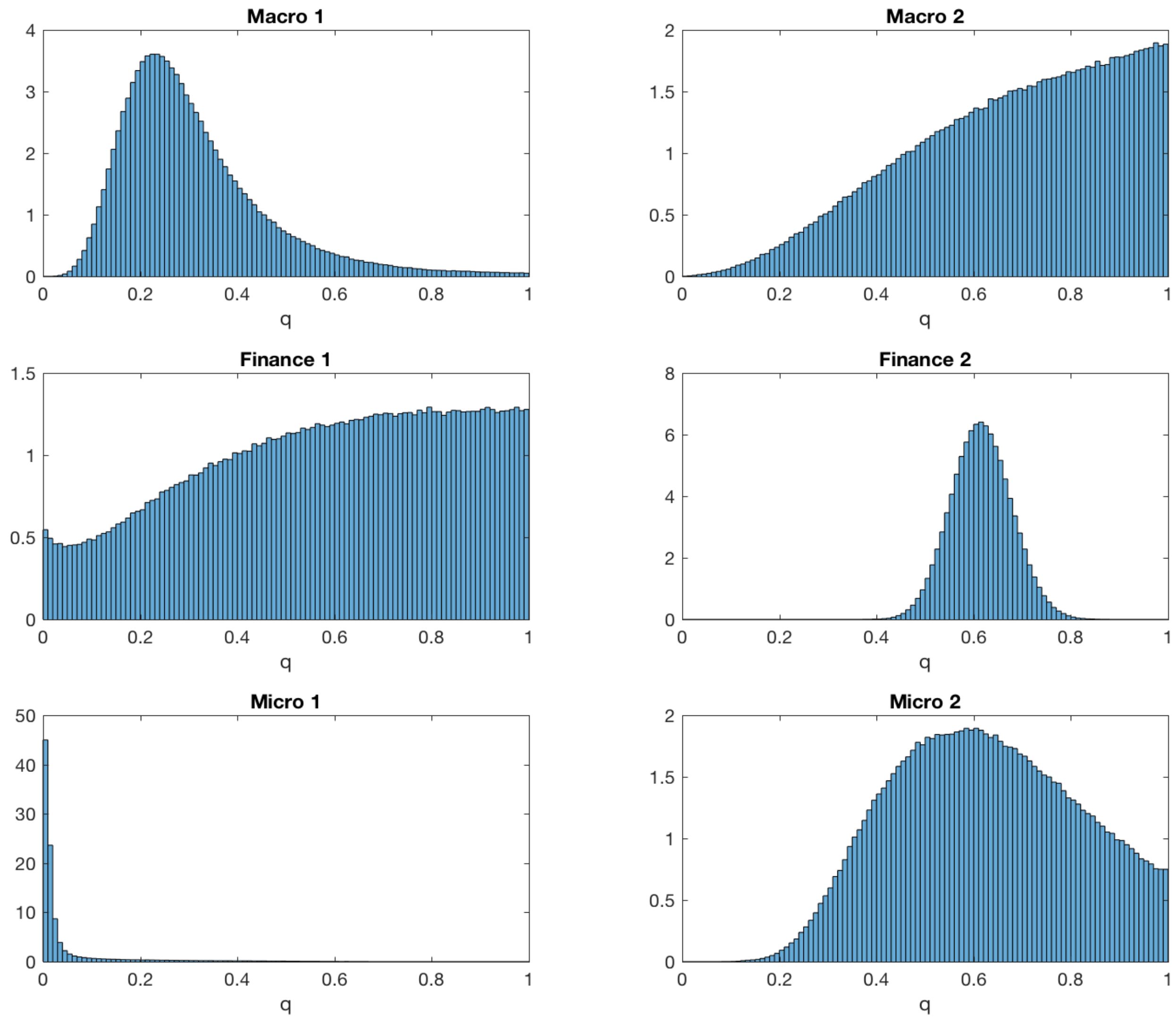
Probability of inclusion and prior variance (q and γ^2)



Probability of inclusion and prior variance (q and γ^2)



Posterior probability of inclusion: $\mathrm{p}(q|Y)$



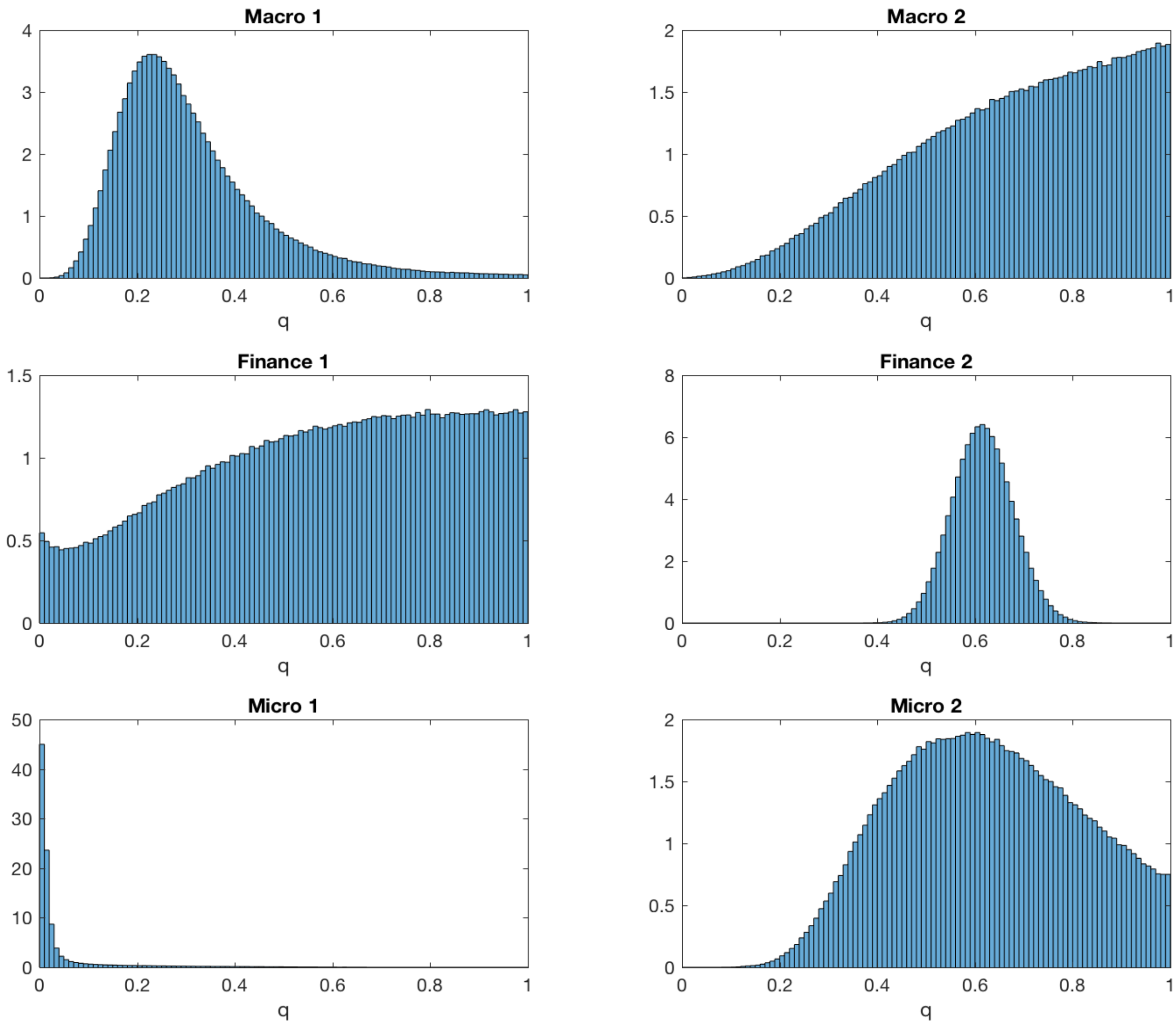
Exploring the posterior

0. Inclusion probability and shrinkage are complements, but imperfect

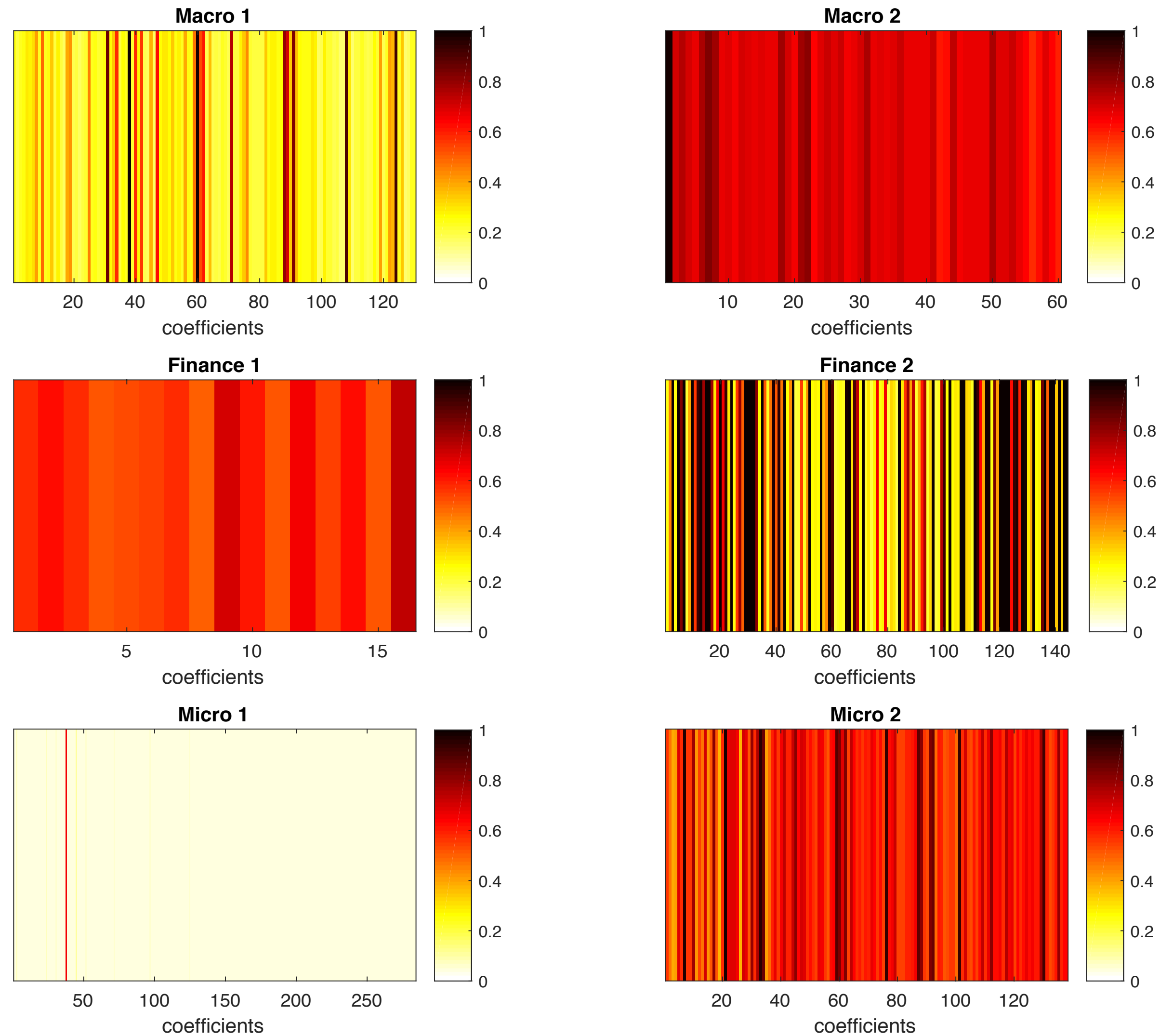
1. No clear pattern of sparsity
   - Posterior not concentrated on a single sparse model, but on a wide set

2. More sparsity emerges only if very tight prior favoring small models

Posterior probability of inclusion: $p(q|Y)$



Patterns of sparsity:
Probability of inclusion of each coefficient



Exploring the posterior

0. Inclusion probability and shrinkage are complements, but imperfect

1. No clear pattern of sparsity
   - Posterior not concentrated on a single sparse model, but on a wide set
     - Predictors rarely systematically excluded
     - Model uncertainty is pervasive
     - Best predictions not with single model, but mixture of many (BMA)

2. More sparsity emerges only if very tight prior favoring small models

Exploring the posterior

0. Inclusion probability and shrinkage are complements, but imperfect

1. No clear pattern of sparsity
   - Posterior not concentrated on a single sparse model, but on a wide set

2. More sparsity emerges only if very tight prior favoring small models

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t\ ,\ \ \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} \mathcal{N}(0, \sigma^2 \gamma^2) & \text{with probability} \quad q \\ 0 & \text{with probability} \quad 1 - q \end{cases}$$

- Hyperpriors

$$q \sim \mathcal{B}(1, 1), \qquad \frac{q\ k\ var(x)\ \gamma^2}{q\ k\ var(x)\ \gamma^2 + 1} \equiv R^2 \sim \mathcal{B}(1, 1)$$

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t \,, \;\; \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} \mathcal{N}(0, \sigma^2 \gamma^2) & \text{with probability} \quad q \\ 0 & \text{with probability} \quad 1 - q \end{cases}$$

- Hyperpriors

$$q \sim \mathcal{B}(1, k), \qquad \frac{q\; k\; var(x)\; \gamma^2}{q\; k\; var(x)\; \gamma^2 + 1} \equiv R^2 \sim \mathcal{B}(1, 1)$$

The predictive model

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t \,, \;\; \varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Prior

$$\beta_i | \sigma^2, \gamma^2, q \underset{iid}{\sim} \begin{cases} \mathcal{N}(0, \sigma^2 \gamma^2) & \text{with probability} \quad q \\ 0 & \text{with probability} \quad 1-q \end{cases}$$

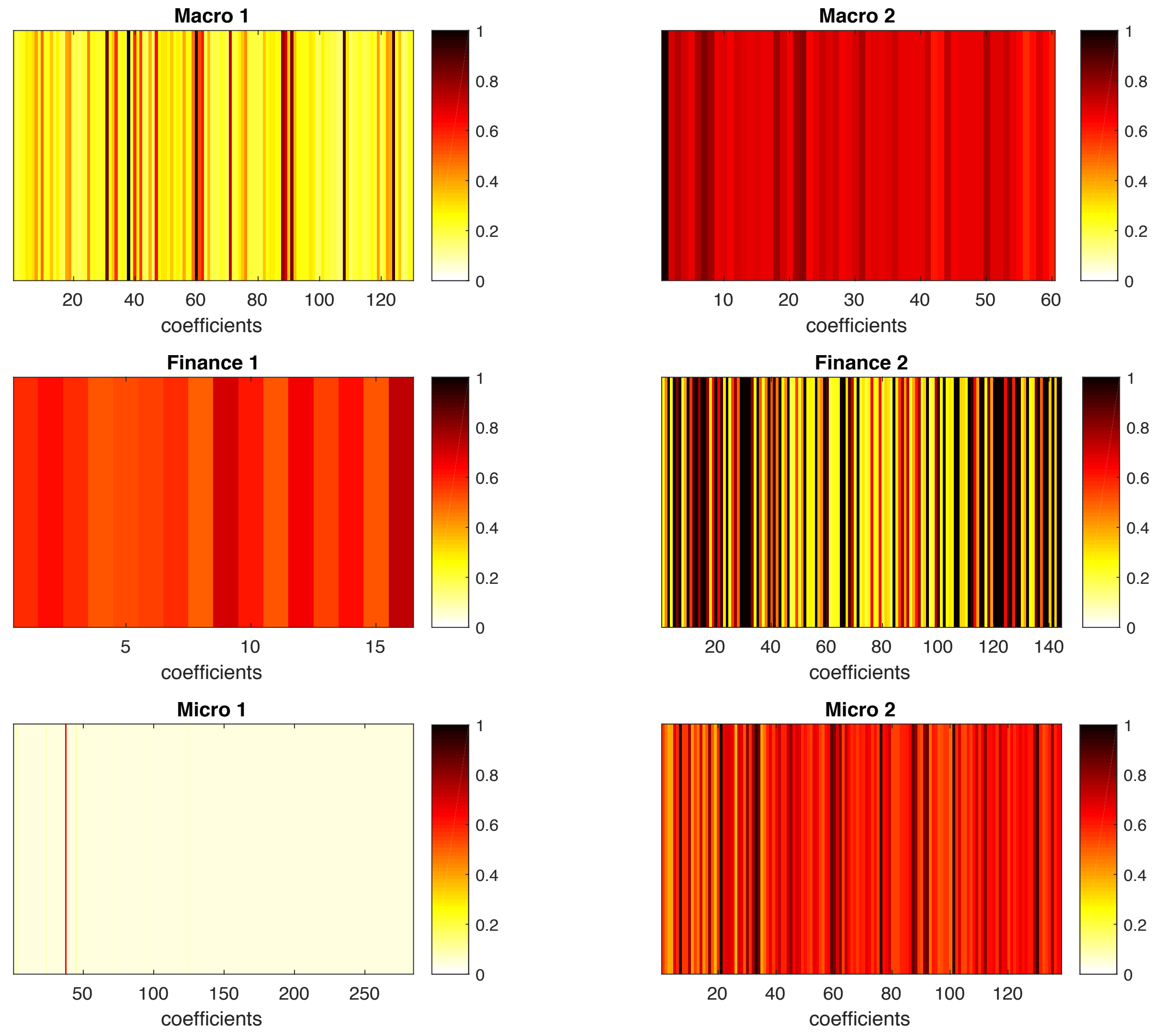
Castillo et al. (2015, Annals)

- Hyperpriors

$$q \sim \mathcal{B}(1, k), \qquad \frac{q\, k\, var(x)\, \gamma^2}{q\, k\, var(x)\, \gamma^2 + 1} \equiv R^2 \sim \mathcal{B}(1, 1)$$

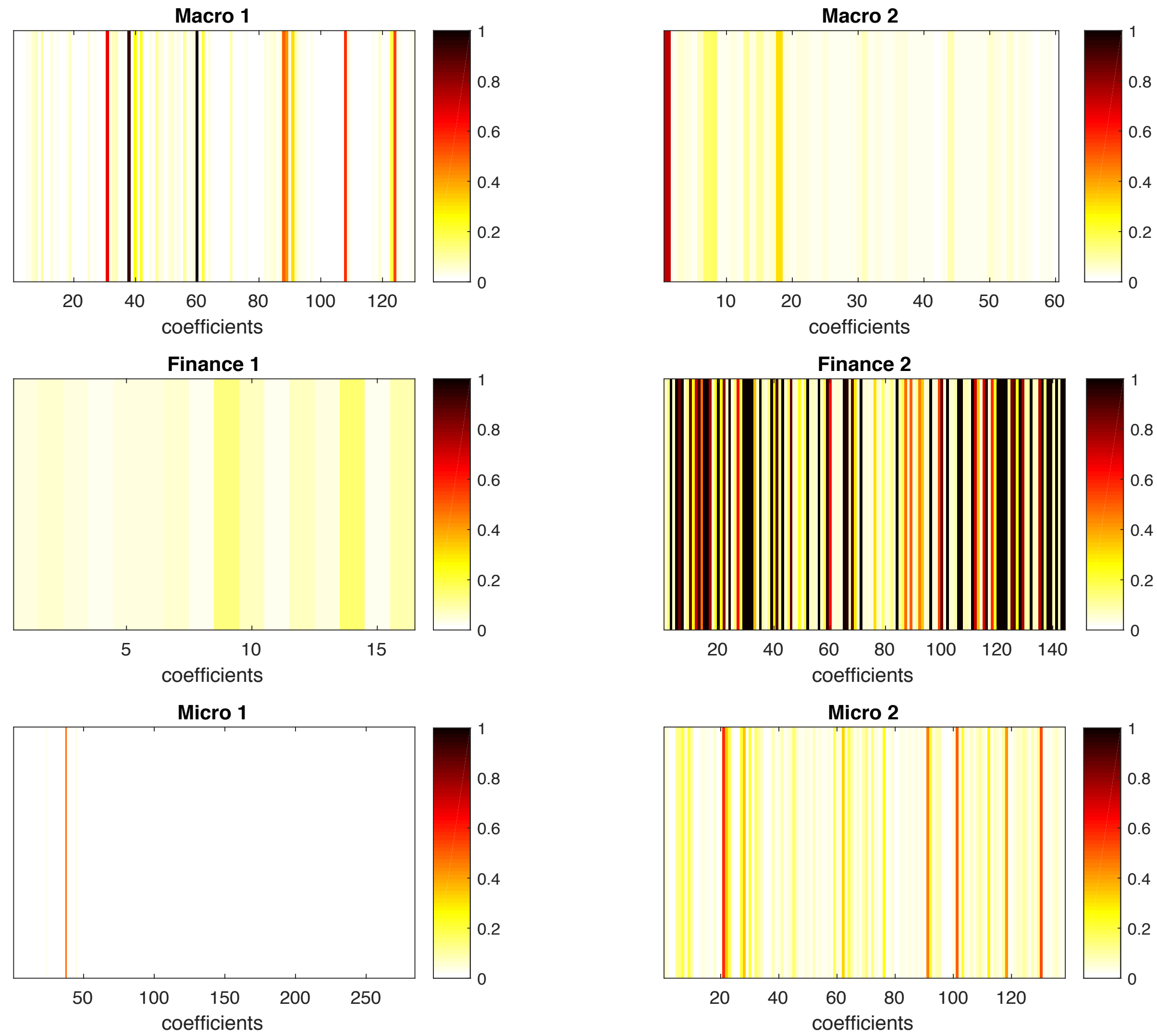
Patterns of sparsity with a flat prior on q

Probability of inclusion of each coefficient



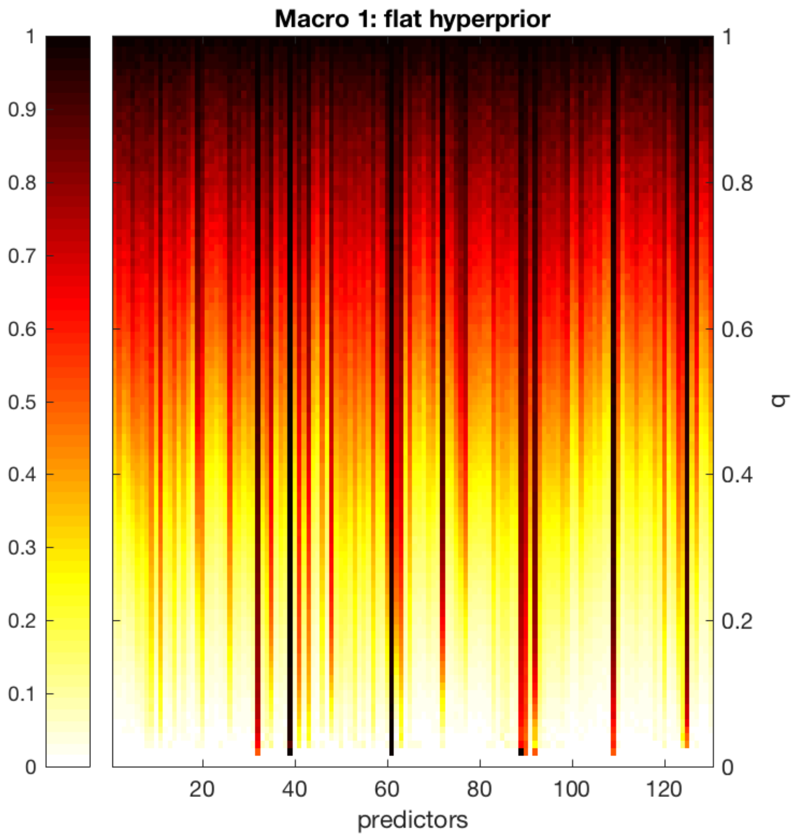
Patterns of sparsity with a tight prior on low q

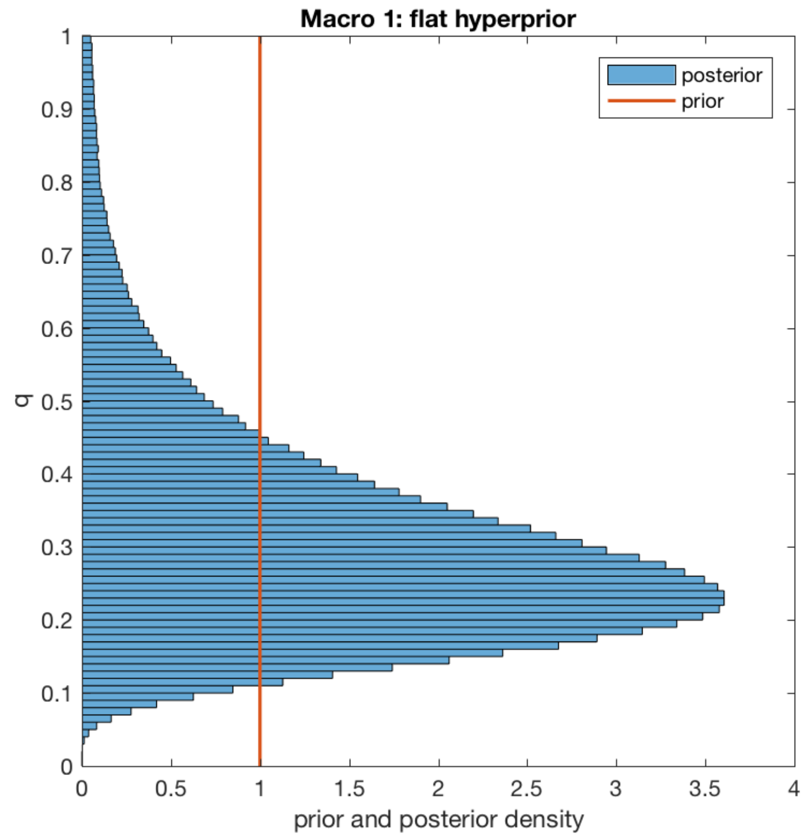
Probability of inclusion of each coefficient



Patterns of sparsity with a flat prior on q

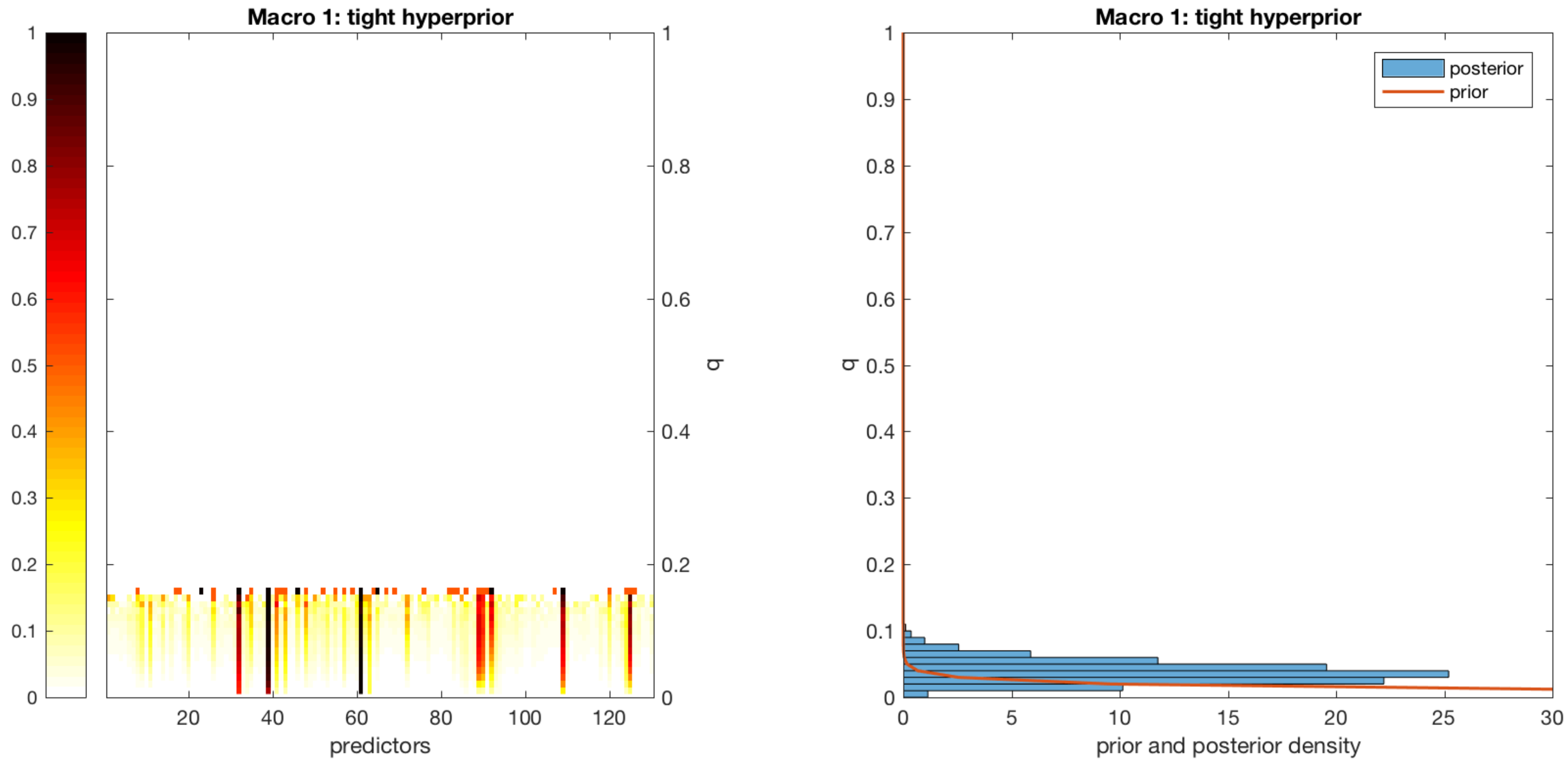
Probability of inclusion of each coefficient, given q



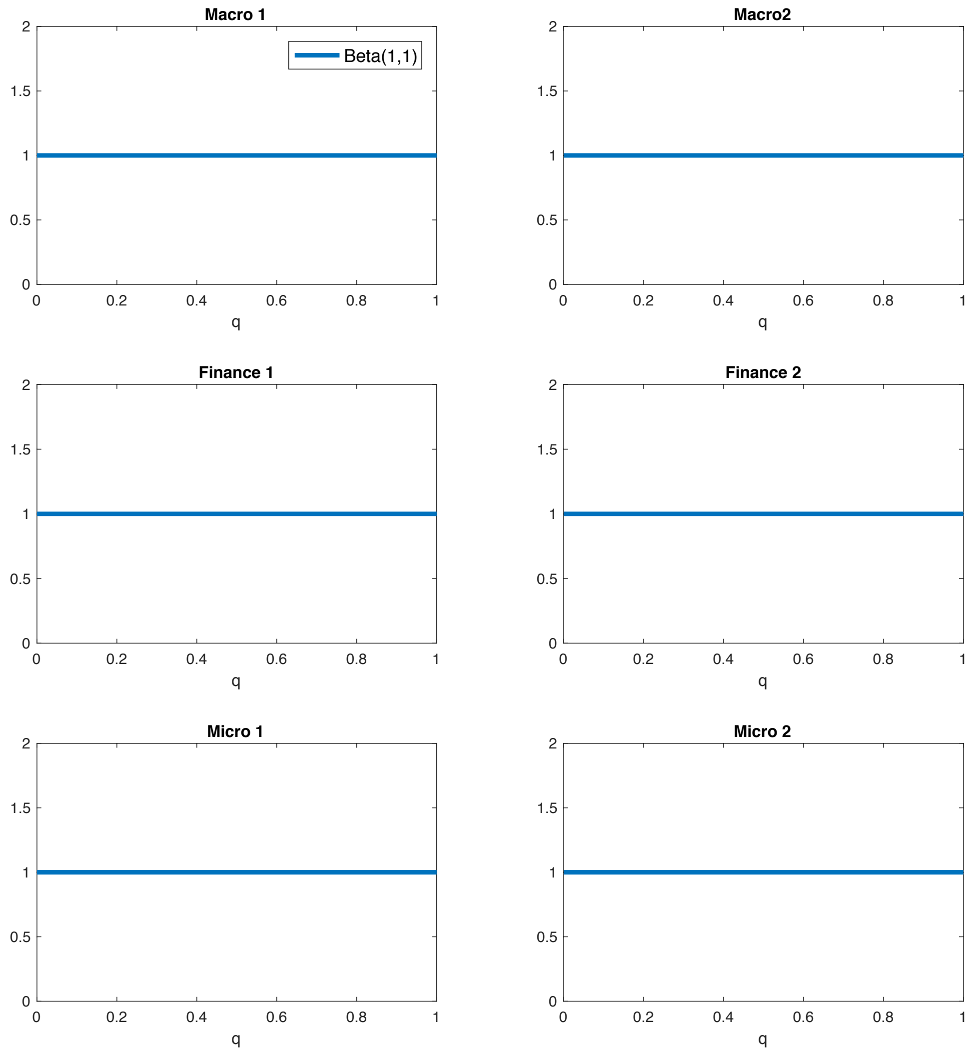


Patterns of sparsity with a tight prior on q

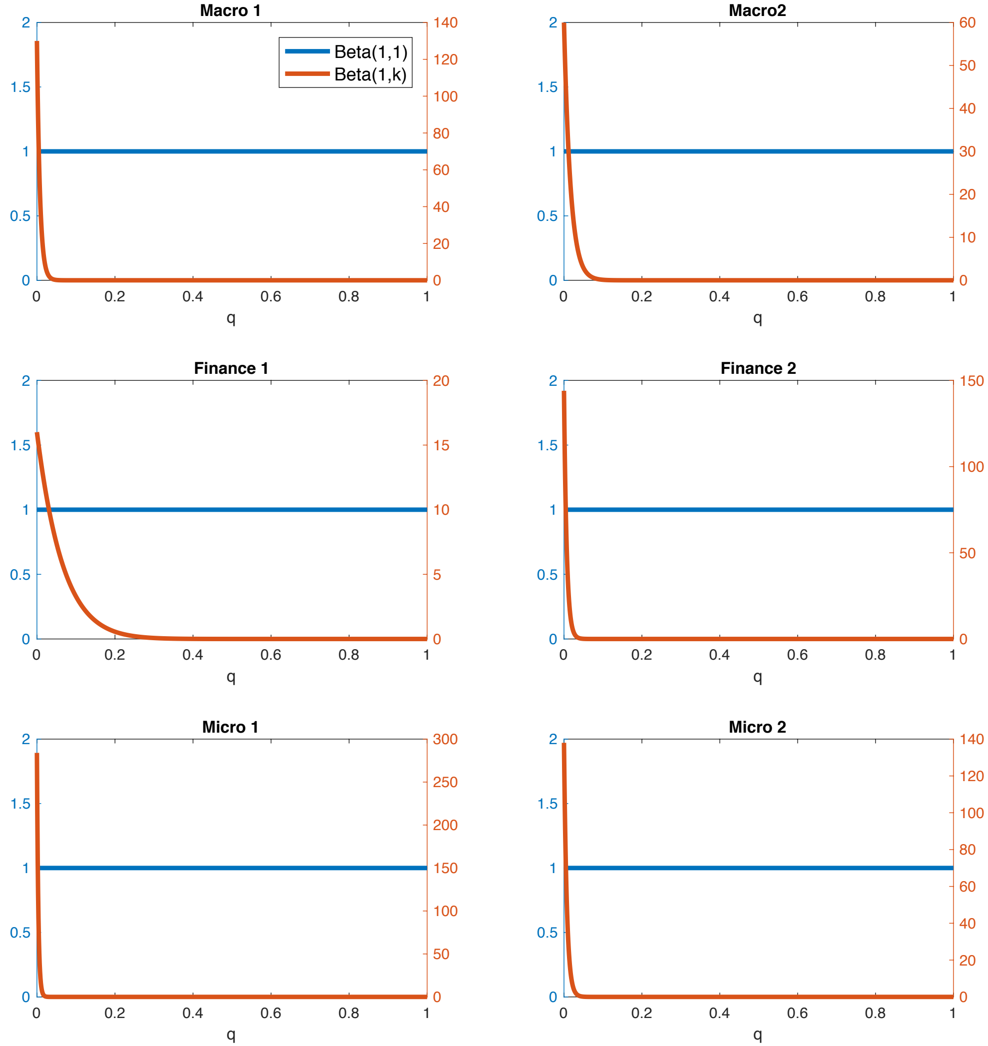
Probability of inclusion of each coefficient, given q



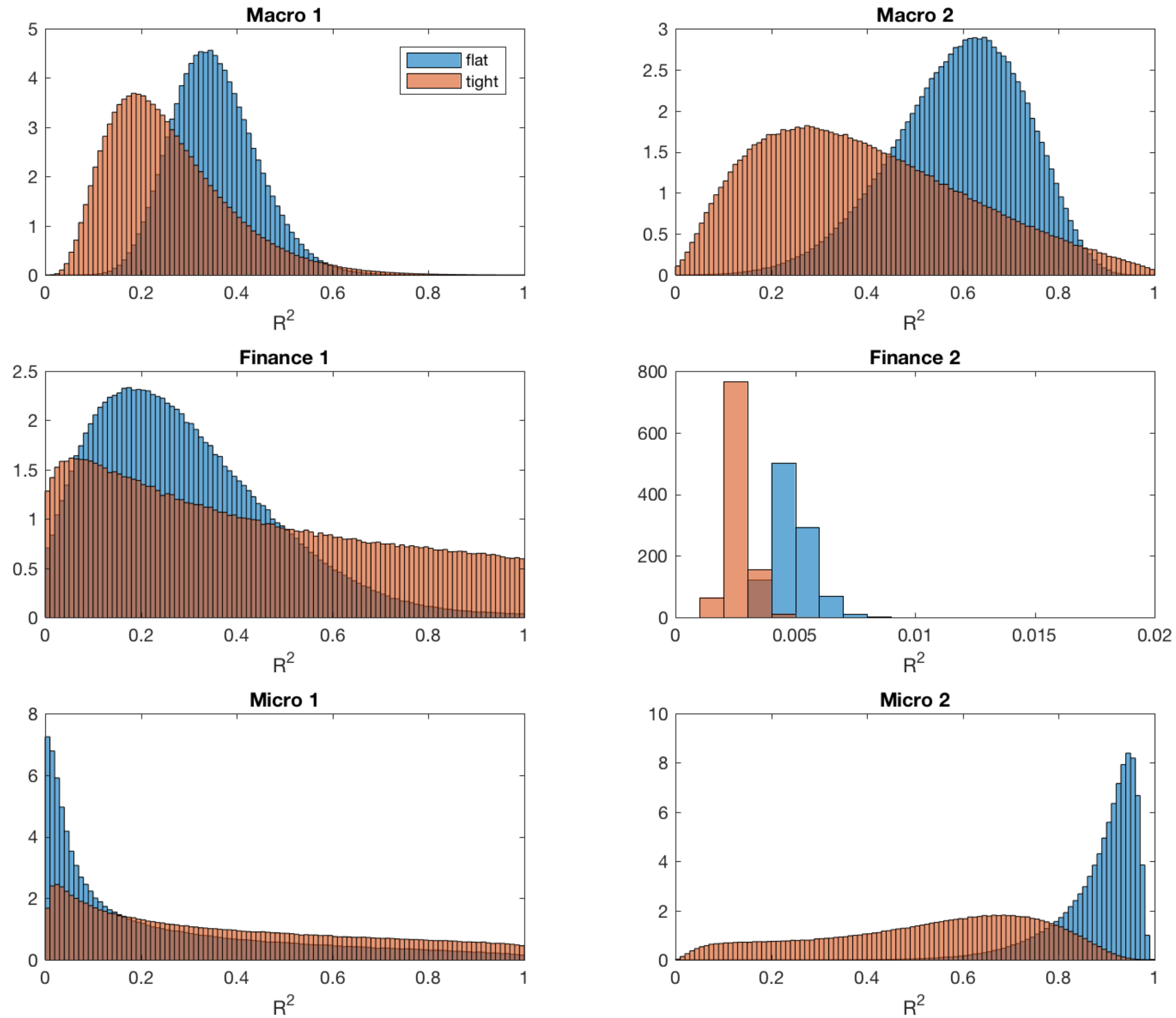
Baseline hyperprior: flat on q



Alternative hyperprior: tight on low q



Posterior of R^2 with a *flat* and a *tight* prior on q



Summing up

0. Inclusion probability and shrinkage are complements, but imperfect

1. No clear pattern of sparsity
   - Posterior not concentrated on a single sparse model, but on a wide set

2. More sparsity emerges only if very tight prior favoring small models

The illusion of sparsity

The prior distribution

- Alternative representation

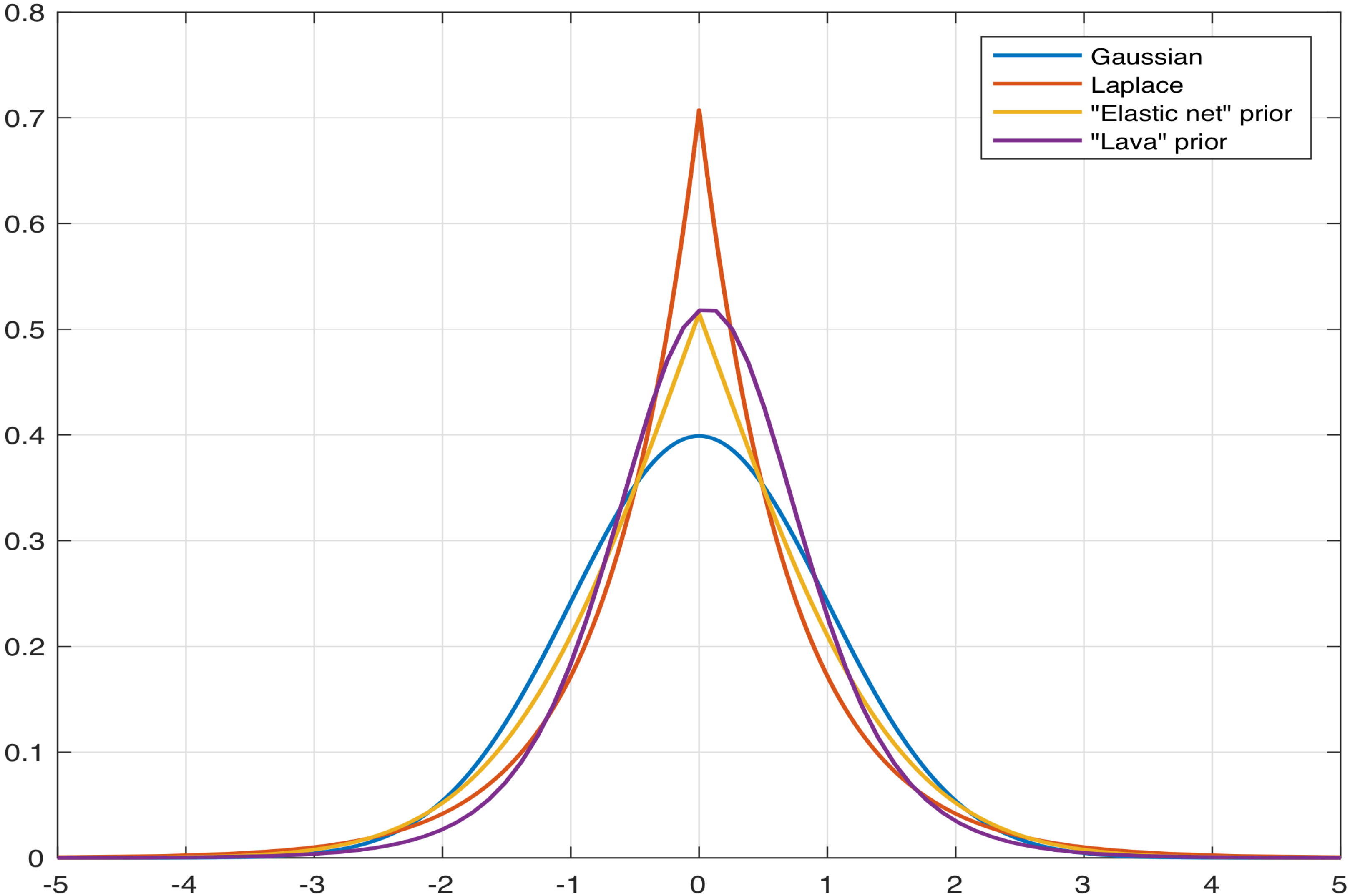
$$\beta_i|\sigma^2,\gamma^2,q \underset{iid}{\sim} \mathcal{N}(0,\sigma^2\gamma^2 z_i),$$

$$z_i|q \underset{iid}{\sim} Bernoulli(q)$$

- Relation with other popular shrinkage methods
  - Ridge: $q = 1$
  - Lasso: $z_i \underset{iid}{\sim} \mathcal{E}xponential$
  - Lava: $z_i \underset{iid}{\sim} Shifted\ \mathcal{E}xponential$
  - Horse shoe: $z_i \underset{iid}{\sim} Half\ Cauchy$
  - Elastic net: $z_i \underset{iid}{\sim} transformation\ of\ a\ truncated\ Gamma$

- None admits a sparse representation of with positive probability

Bayesian interpretation of various shrinkage methods



$p(q|Y)$ as a measure of predictive accuracy

- Posterior of q

$$p(q|Y) \propto p(Y|q) \cdot p(q)$$

$p(q|Y)$ as a measure of predictive accuracy

- Posterior of *q*

$$p(q|Y) \propto p(Y|q)$$

$p(q|Y)$ as a measure of predictive accuracy

- Posterior of q

$$p(q|Y) \propto p(Y|q) = \prod_{t}^{T} p(y_t|y^{t-1}, q)$$

$p(q|Y)$ as a measure of predictive accuracy

- Posterior of *q*

$$p(q|Y) \propto p(Y|q) = \prod_{t}^{T} p(y_t|y^{t-1}, q)$$

predictive score

$p(q|Y)$ as a measure of predictive accuracy

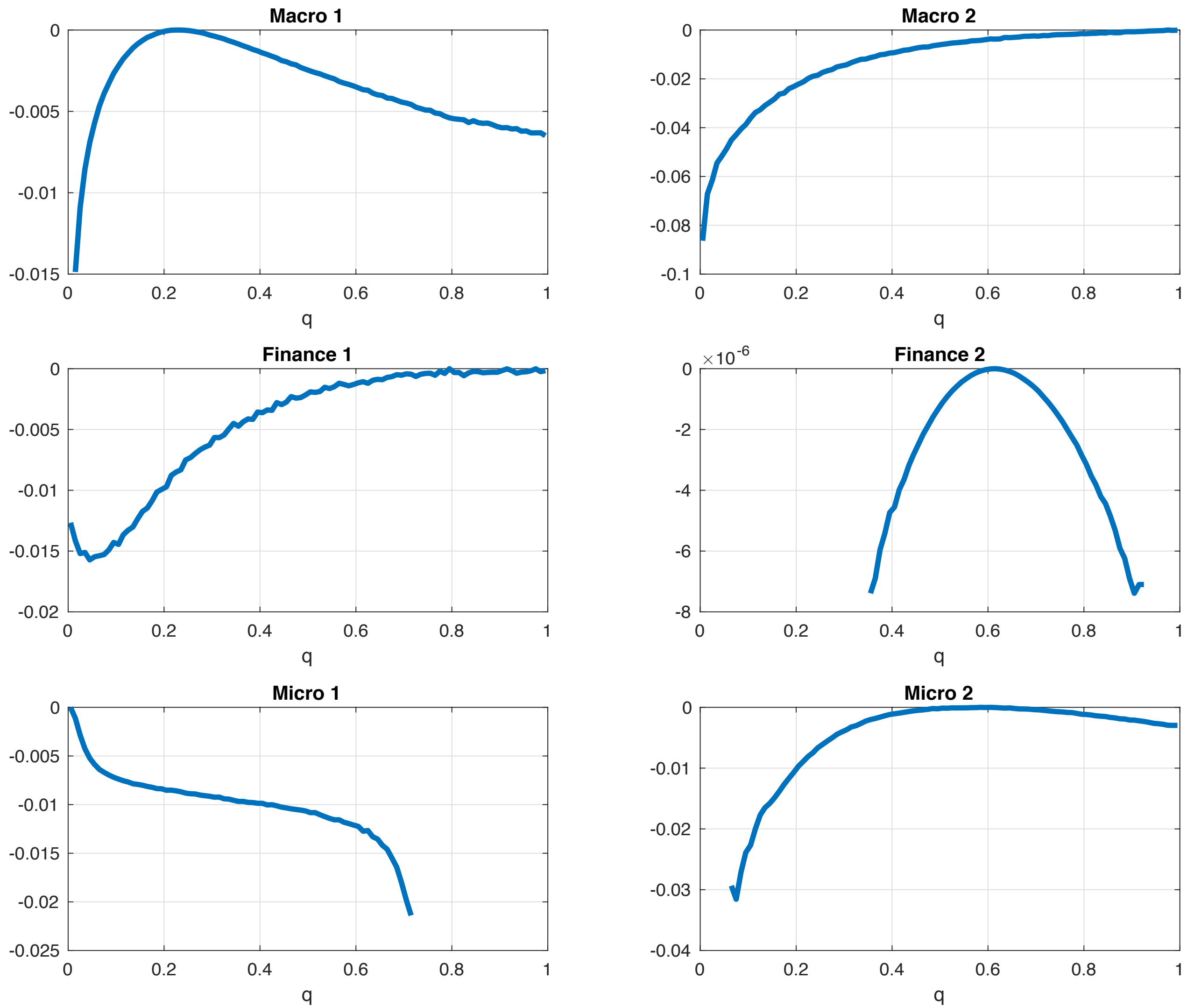
- Posterior of q

$$p(q|Y) \propto p(Y|q) = \prod_{t}^{T} p(y_t|y^{t-1}, q)$$

➥ Average log-predictive score

$$\frac{1}{T}\sum_{t}^{T} \log p(y_t|y^{t-1}, q) = \frac{1}{T} log\, p(q|Y) + constant$$

Average log-predictive score, relative to best fitting model



Probability of inclusion (q) and R^2