Measuring Prior Sensitivity and Prior Informativeness in Large Bayesian Models

Ulrich K. Müller
Princeton University

January 2012

Introduction

- Bayesian estimation of models with many parameters has become standard tool in empirical macroeconomics
- Prior matters unless data very informative
- Difficult to assess role of prior and likelihood when there are many parameters
- Standard practice: Compare marginals of prior and posterior distribution

Bivariate Example

- Observe two Gaussian RVs

$$\begin{aligned} Y_1 &= \theta_1 + 10\varepsilon_1 + \varepsilon_2/10 \\ Y_2 &= \theta_2 + 10\varepsilon_1 - \varepsilon_2/10 \end{aligned}$$

  where $\varepsilon_1, \varepsilon_2 \sim iid\mathcal{N}(0,1)$.

- Interest is exclusively on θ_1

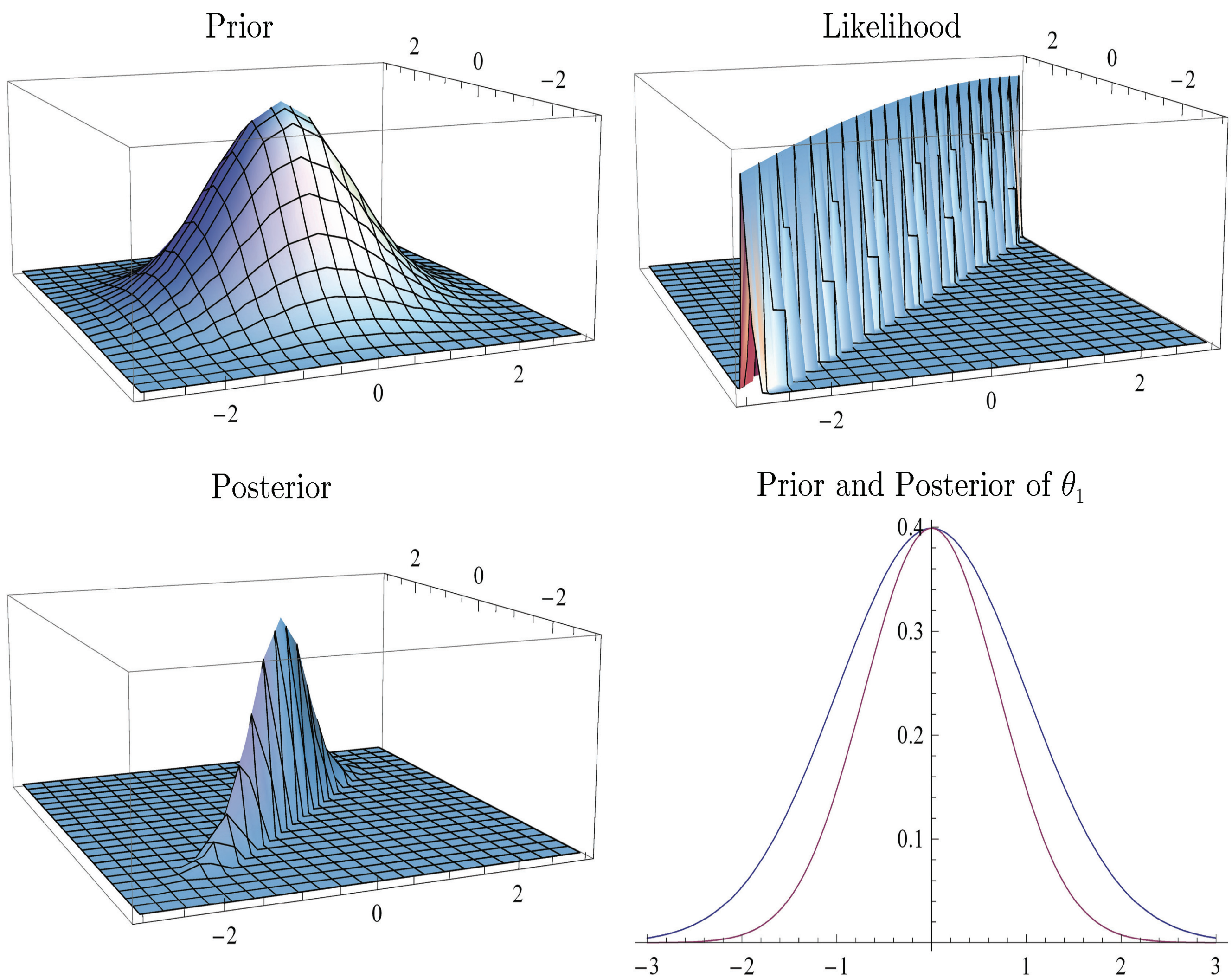
- Without knowledge of θ_2, only Y_1 is informative about θ_1

- Since $Y_1 \sim \mathcal{N}(\theta_1, 100.01)$, very little information about θ_1 in likelihood: With prior $\theta_1 \sim \mathcal{N}(0,1)$, $\theta_1|Y_1 = 0 \sim \mathcal{N}(0, 0.990)$

- Yet full Bayesian analysis with prior $\theta \sim \mathcal{N}(0, I_2)$:

$$\theta_1|Y = 0 \sim \mathcal{N}(0, 0.504)$$

Bivariate Example



Overview I

- Develop statistics that help to answer two questions: Given a scalar parameter of interest

  1. How sensitive are posterior results to variations in the prior?

  2. How informative is prior relative to likelihood?

- Basic idea: Study variation of posterior mean as a function of prior mean for both questions

  – If likelihood is flat, posterior is like prior, and prior mean changes are pushed through to the posterior one-to-one. Indicates both prior sensitivity and strong (relative) prior informativeness.

  – If likelihood is very peaked, posterior largely unaffected by prior changes. Indicates both prior robustness and low prior informativeness.

- Implementation via local prior mean changes, that is study of derivative of posterior mean with respect to prior mean.

Overview II

- Prior mean change via exponential family embedding

  $\Rightarrow$ Derivative matrix becomes simple function of prior and posterior covariance matrices, easily computed from MCMC output

- Prior sensitivity measure **PS** is Euclidian norm of (normalized) derivative vector: measures maximal change of posterior mean by varying prior mean by the multivariate analogue of one prior standard deviation

- Prior informativeness **PI** $\in [0, 1]$ measures fraction of prior information for posterior results
  - **PI** equal to derivative in scalar parameter case
  - **PI** derived from derivative matrix via axiomatic requirements in vector parameter case

Related Literature

- Bayesian local sensitivity analysis. In particular, local sensitivity of posterior mean with respect to parametric change in prior: Basu, Jammalamadaka, and Liu (1996) and Perez, Martin, and Rufo (2006)

  Contribution regarding **PS** merely exponential family embedding, and normalization

- No close counterpart to **PI**

  Recent literature that studies identification of DSGE models: Canova and Sala (2009), Iskrev (2010a, 2010b) and Komunjer and Ng (2009)

  - **PI** not binary "identification or not", but measures relative importance of prior
  - **PI** not tied to linear Gaussian framework
  - **PI** not based on frequentist identification concept, but likelihood based

Model with Scalar Parameter

- θ is scalar, p prior density with $\mu_p = E_p[\theta]$ and $\sigma_p^2 = V_p[\theta]$, π is posterior density under prior p with $\sigma_\pi^2 = V_\pi[\theta$

- Embed p in family p_α indexed by α

$$p_\alpha(\theta) = C(\alpha)\exp\left[\frac{\alpha(\theta - \mu_p)}{\sigma_p^2}\right] p(\theta)$$

  so that for α small, $E_{p_\alpha}[\theta] \approx E_p[\theta] + \alpha$

- Derivative of posterior mean $\mu_\pi(\alpha)$ with respect to prior mean

$$\frac{d\mu_\pi(\alpha)}{d\alpha}|_{\alpha=0} = J = \sigma_\pi^2/\sigma_p^2$$

- PS $= \sigma_p J$: linear approximation to change in posterior mean when prior mean is increased by one prior standard deviation

- PI $= \min(J, 1)$: "push-through" rate of prior mean change to posterior mean change

PI as Fraction of Prior Information

- Suppose prior log-density and log-likelihood are quadratic in θ, i.e. $p_\alpha(\theta) \propto \exp[-\frac{1}{2}\frac{(\theta-\mu_p-\alpha)^2}{\sigma_p^2}]$ and $l(\theta) \propto \exp[-\frac{1}{2}\frac{(\theta-\mu_l)^2}{\sigma_l^2}]$.

- By standard calculation, $\sigma_\pi^{-2} = \sigma_p^{-2} + \sigma_l^{-2}$ and

$$\mu_\pi(\alpha) = w(\mu_p + \alpha) + (1-w)\mu_l \quad \text{with} \quad w = \frac{\sigma_p^{-2}}{\sigma_p^{-2} + \sigma_l^{-2}}$$

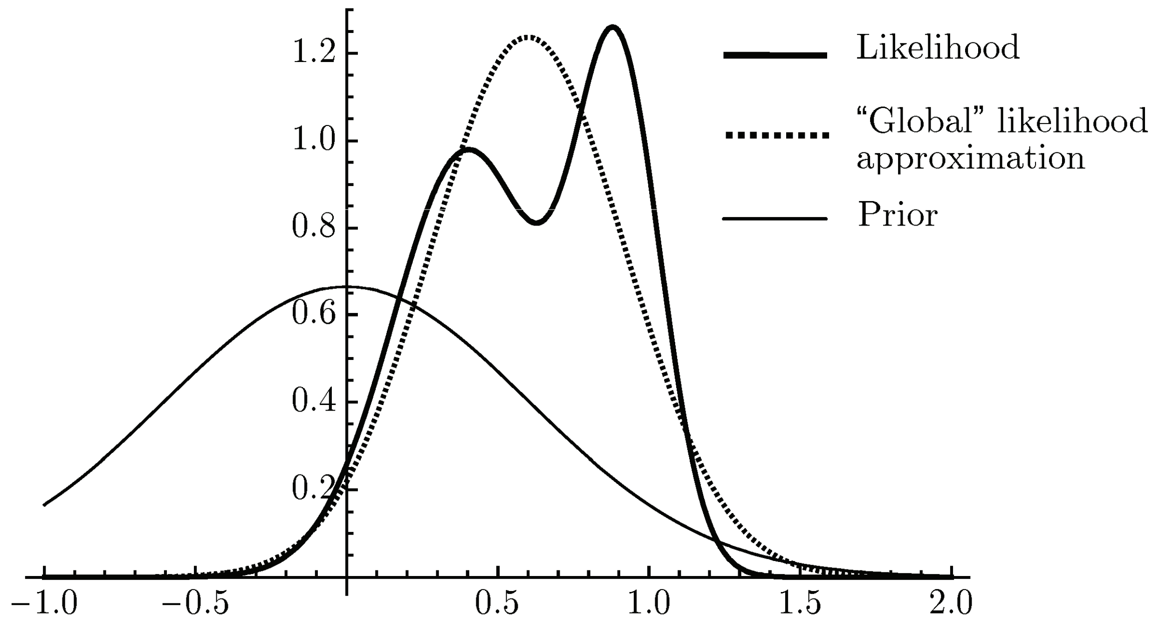
so that

$$\text{PI} = \frac{d\mu_\pi(\alpha)}{d\alpha}|_{\alpha=0} = w$$

is ratio of prior information σ_p^{-2} to total information $\sigma_p^{-2} + \sigma_l^{-2}$.

- Interpretation remains reasonable approximation if prior log-density and log-likelihood are only approximately quadratic

Example



- Likelihood of a mixture of two normals Y with $E[Y] = \theta$

- "Global" quadratic log-likelihood approximation with μ_l and σ_l^2 computed from scale-normalized likelihood

- $w = \frac{\sigma_p^{-2}}{\sigma_p^{-2} + \sigma_l^{-2}}$ yields $w = 0.224$, and PI $= 0.249$

Model with Vector Parameter

- θ is $k \times 1$ vector, with prior variance $V_p[\theta]$ and baseline posterior variance $V_\pi[\theta]$.

- Embed prior p in exponential family

$$p_\alpha(\theta) = C(\alpha)\,\mathsf{exp}[\alpha' V_p[\theta]^{-1}(\theta - \mu_p)]p(\theta)$$

  so that for small α, $E_{p_\alpha}[\theta] \approx E_p[\theta] + \alpha$

- $k \times k$ derivative matrix

$$J = \frac{\partial \mu_\pi(\alpha)}{\partial \alpha'}|_{\alpha=0} = V_\pi[\theta] V_p[\theta]^{-1}$$

PS **with Vector Parameter**

- $v'\theta$ is scalar parameter on interest
- Derivative vector of the posterior mean of $v'\theta$ is $v'J$
- Define

$$\text{PS} = \max_{\alpha' V_p[\theta]^{-1}\alpha = 1} v' J\alpha = \sqrt{v' V_\pi[\theta] V_p[\theta]^{-1} V_\pi[\theta] v}$$

$\Rightarrow$ largest change of the posterior mean of θ that can be induced by multivariate analogue of "one standard deviation change" of prior mean

PI with Vector Parameter: Gaussian Case

- Suppose $Y \sim \mathcal{N}(\theta, \Sigma)$ with Σ known, and prior $\theta \sim \mathcal{N}(\mu_p, V_p[\theta])$. Parameter of interest is $v'\theta$.

- Without knowledge of θ, likelihood information about $v'\theta$ is summarized by scalar random variable $v'Y \sim \mathcal{N}(v'\theta, v'\Sigma v)$, and prior on $v'\theta$ is $\mathcal{N}(v'\mu_p, v'V_p[\theta]v)$.

  $\Rightarrow$ Fraction of information formula yields

$$
\begin{aligned}
\mathsf{PI}_G &= \frac{(v'V_p[\theta]v)^{-1}}{(v'V_p[\theta]v)^{-1} + (v'\Sigma v)^{-1}} \\
&= 1 - \frac{v'V_p[\theta]v}{v'V_p[\theta](V_p[\theta] - V_\pi[\theta])^{-1}V_p[\theta]v}
\end{aligned}
$$

- Bivariate example of Introduction: $\mathsf{PI}_G = 0.990$.

PI with Vector Parameter: Axiomatic Approach

- Gaussian case special.
- In general, potential prior informativeness measures **PI** based on (normalized) derivative matrix $J \in \mathbb{R}^{k \times k}$ can be thought of as mappings $\mathsf{PI} : \mathbb{R}^{k \times k} \mapsto [0, 1]$.
- Impose axiomatic requirements on such mappings that make sense for a prior informativeness measure.
- Paper identifies a set of "reasonable" requirements that imply

$$\mathsf{PI} = \begin{cases} 1 & \text{if } \lambda_{\mathsf{max}}(J) \geq 1 \\ \mathsf{PI}_G & \text{otherwise} \end{cases}$$

$\Rightarrow$ **PI** interesting statistic also outside Gaussian case

Relationship to Frequentist Identification

- Rothenberg (1971) defines $\theta_0 \in \Theta$ to be *identifiable* if $f(y;\theta) = f(y;\theta_0)$ for all $y \in \mathcal{Y}$ implies $\theta = \theta_0$.

- Entirely flat $l(\theta) = f(y;\theta)$ for observed $Y = y$ not incompatible with identifiability, as other draws of Y might have been informative. But with $l(\theta)$ flat, observed data not at all informative, and PI $= 1$ correctly communicates that.

- If density is constant only over "small" set Θ', then lack of identifiability, but also lack of useful information about θ? PI continues to summarize global shape of likelihood.

- PI not binary, and measures data informativeness about *parameter* $v'\theta$, not identification at a particular parameter *value* θ_0.

Conditional PI Analysis

- Interest is in θ_j. Suppose known that data not informative about θ_i, so prior on θ_i is important. Is prior on parameters other than θ_i important for posterior of θ_j?

- Perform analysis conditional on prior about θ_i by dropping ith row and column of $V_p[\theta]$ and $V_\pi[\theta]$ in computation of PI for θ_j

- Justification in two stage information acquisition about θ_i

  1. Previous study A updates very vague prior $p_{A,i}$ on θ_i with variance $\sigma^2_{A,p,i}$ to tighter posterior with variance $\sigma^2_{A,\pi,i}$
  2. Current study B uses posterior on θ_i as prior, $\sigma^2_{p,i} = \sigma^2_{B,p,i} = \sigma^2_{A,\pi,i}$

  With no further links between studies, current posterior is also posterior for combined data set with prior $p_{A,i}$ on θ_i

  As $\sigma^2_{A,p,i} \rightarrow \infty$, PI above is prior informativeness relative to combined data set

Application to Smets and Wouters (2007)

- DSGE model with sticky prices and wages, habit formation in consumption, variable capital utilization and investment adjustment costs
- 14 endogenous variables, driven by 7 exogenous shocks described by 17 parameters (="shock parameters")
- 24 additional parameters describing the model, of which SW fix 5, so that 19 parameters are estimated ("=structural parameters")
- (Essentially) same prior as in SW

Results

- $\lambda_{\text{max}} = 1.25$, with eigenvector loading of 0.95 on $\bar{\pi}$

  $\Rightarrow$ PI conditional on prior information about $\bar{\pi}$.

- Selected parameters:

		Prior		Posterior				
		μ_p	σ_p	μ_π	σ_π	σ^2_π/σ^2_p	PI	PS
φ	$\mathcal{N}$	4.00	1.50	5.74	1.03	0.48	0.53	0.75
ι_p	$\mathcal{B}$	0.50	0.15	0.25	0.09	0.35	0.37	0.06
ξ_w	$\mathcal{B}$	0.50	0.10	0.70	0.07	0.43	0.75	0.06
ξ_p	$\mathcal{B}$	0.50	0.10	0.65	0.06	0.31	0.50	0.04
ω_p	$\mathcal{IG}$	0.30	0.20	0.15	0.02	0.01	0.01	0.01
ρ_p	$\mathcal{B}$	0.50	0.20	0.89	0.05	0.06	0.10	0.03

- Shock parameters have low PI (well pinned-down by likelihood), and so do Impulse Responses and Variance Decompositions

  $\Rightarrow$ structural parameter have limited role for determination IRs and VDCs

Conclusion

- Suggestion of two statistics that shed light on role of prior in models with high dimensional parameters
  - **PS** measures sensitivity of posterior mean of parameter of interest to variation in prior means
  - **PI** quantifies to which degree the posterior results for parameter of interest are driven by prior information
- Entirely straightforward to implement with MCMC output
- Potentially other useful statistics. But some reasonable requirements lead to suggested measures