Bayesian Econometrics

April 6, 2013

Luc Bauwens

CORE

Université catholique de Louvain

COURSE STRUCTURE

- Chapter 1: Concepts
- Chapter 2: Numerical Methods (p 33)
- Chapter 3: Single Equation Regression Analysis (p 95)
- Chapter 4: VAR Models (p 149)

CHAPTER 1

- 1.1 Bayesian inference
- 1.2 Criteria for evaluating statistical procedures
- 1.3 Probability: objective or subjective?
- 1.4 Skills and readings

Really?

- "The Bayesian approach to econometrics is conceptually simple and, following recent developments computationally straightforward."

  Tony Lancaster (2004)

  An Introduction to Modern Bayesian Econometrics. Balckwell, page 58.

Bayesian inference

- An approach to statistical inference on model parameters, quite different from classical methods (like ML, GMM, ...).
- The differences concern:

  1-the treatment of parameters of models (random variables versus fixed constants);

  2-the criteria for evaluating statistical procedures of estimation and hypothesis testing (conditional only on observed data versus in terms of sampling properties);

  3-the interpretation of probability (subjective versus objective).

Principle of Bayesian inference

- Bayesian inference formally treats the unknown parameters of a model as random variables.
- The state of knowledge about the parameters may be represented through a probability distribution, the prior distribution (‘prior’ to observing data).
- Data supposed to be generated by the model provide information about the parameters. The data information is available through the data density that is the likelihood function when considered as a function of the parameters.
- Prior information and data-based information are combined through Bayes theorem and provide a posterior distribution.

Bayes theorem

- Formally, Bayes theorem (data $y \in S$ and parameters $\theta \in \Theta$) provides the posterior density (assuming θ is continuous) from the prior density $\varphi(\theta)$ and the data density $f(y|\theta)$:

$$\varphi(\theta|y) = \frac{\varphi(\theta) f(y|\theta)}{f(y)}.$$

- $f(y)$ is the marginal density of y, also called predictive density:

$$f(y) = \int \varphi(\theta) f(y|\theta) d\theta.$$

Sequential sampling

- Bayes theorem provides a coherent learning process: from prior to posterior.
- Learning may be progressive. Suppose that y_1 is a first sample and y_2 a second one, i.e. $y_1 \sim f(y_1|\theta)$ and $y_2 \sim f(y_2|y_1, \theta)$.
- The posterior after observing the first sample is

$$\varphi(\theta|y_1) = \frac{\varphi(\theta) f(y_1|\theta)}{f(y_1)}.$$

where $f(y_1) = \int \varphi(\theta) f(y_1|\theta) d\theta$. This posterior serves as the prior for using the second sample.

Sequential sampling

- The updated posterior based on y_2 is

$$\varphi(\theta|y_2, y_1) = \frac{\varphi(\theta|y_1) f(y_2|y_1,\theta)}{f(y_2|y_1)} = \frac{\varphi(\theta) f(y_1|\theta) f(y_2|y_1,\theta)}{f(y_1) f(y_2|y_1)}$$

  where $f(y_2|y_1) = \int \varphi(\theta|y_1) f(y_2|y_1,\theta) d\theta$.

- This posterior is the same as the one obtained by applying Bayes theorem to the joint sample $y = (y_1, y_2)$ since

  - $f(y|\theta) = f(y_1|\theta) f(y_2|y_1,\theta)$
  - $f(y) = f(y_1) f(y_2|y_1)$.

Modelling using Bayesian inference

Paraphrasing Lancaster (1, p 9), modelling the Bayesian way may be described in steps:

1. Formulate your economic model as a family of probability distributions $f(y|X, \theta)$ for observable random variables y and X (exogenous variables).
2. Organize your prior beliefs about θ into a prior.
3. Collect the data for y and X, compute and report the posterior.
4. Evaluate your model and revise it if necessary.
5. Use the model for the purposes for which it has been designed (scientific reporting, evaluating a theory, prediction, decision making...).

Summarizing the posterior

Features of the posterior that should be reported are:

- Posterior means and standard deviations. This is a minimum.
- The posterior variance-covariance matrix (or the corresponding correlation matrix).
- Graphs and quantiles of univariate marginal densities of elements of θ that are of particular interest. Skewness and kurtosis coefficients, and the mode(s) if the marginal densities are clearly non-Gaussian.
- Contours of bivariate marginal densities of pairs of elements of θ that are of particular interest.

Simple example

- Sampling process: let $y = (y_1, \ldots, y_n)$ where $y_i|\mu \sim I.N(\mu, 1)$ for $\mu \in R$. The data density is: $f(y|\mu) = (2\pi)^{-n/2} \exp[-0.5 \sum_{i=1}^{n} (y_i - \mu)^2]$.
- Prior density: $\mu \sim N(\mu_0, n_0^{-1})$, where $\mu_0 \in R$ and $n_0 > 0$ are constants chosen to represent prior information: μ_o is a prior idea about the most likely value of μ and n_0 sets the precision (=inverse of variance).
- $\Rightarrow$ Posterior density: $\mu|y \sim N(\mu_*, n_*^{-1})$ where $n_* = n + n_0$ and $\mu_* = \frac{n\bar{y}+n_0\mu_0}{n+n_0}$.

  How do we find that the posterior is normal with expectation μ_* and variance $1/n_*$?

Apply Bayes theorem

- By Bayes theorem: $\varphi(\mu|y) \propto \varphi(\mu) f(y|\mu)$.
  $\propto$ means 'proportional to': here we 'forget' $1/f(y)$ because it is a proportionality constant that does not depend on μ. We can do the same in the prior and the data density:
- $\varphi(\mu|y) \propto \exp[-0.5 n_0 (\mu - \mu_0)^2] \exp[-0.5 n (\mu - \bar{y})^2]$.

  The question to ask at this stage is always: is this a form of density in a known class?

  -If YES: exploit the properties of this class, to get the posterior features you want.

  -If NO: use numerical integration to compute the posterior features.

Calculus

- $\sum_{i=1}^{n}(y_i-\mu)^2 = n(\mu-\bar{y})^2 + \sum_{i=1}^{n}(y_i-\bar{y})^2$.
- Add the arguments of the 2 $\exp$ functions:

$$\begin{aligned}
& n_0(\mu-\mu_0)^2 + n(\mu-\bar{y})^2 \\
&= n_0(\mu^2 - 2\mu_0\mu + \mu_0^2) + n(\mu^2 - 2\bar{y}\mu + \bar{y}^2) \\
&= (n_0+n)\mu^2 - 2(n_0\mu_0 + n\bar{y})\mu + (n_0\mu_0^2 + n\bar{y}^2) \\
&= A\mu^2 - 2B\mu + C \\
&= A(\mu - \tfrac{B}{A})^2 + C - \tfrac{B^2}{A}
\end{aligned}$$

$$\Rightarrow \varphi(\mu|y) \propto \exp[-0.5A(\mu - \tfrac{B}{A})^2] = \exp[-0.5n_*(\mu-\mu_*)^2]$$

Apart from a proportionality constant, this is the normal density $N(\mu_*, n_*^{-1})$.

Comments

- If $n \to \infty$, E$(\mu|y) \to \bar{y}$, Var$(\mu|y) \to 0$, and $\varphi(\mu|y) \to \mathbb{1}_{\{\mu=\bar{y}\}}$ (all mass on $\bar{y}$).
- If $n_0 = 0$, the prior variance is infinite. The prior density has no weight, it is said to be diffuse, or non-informative, or flat. The posterior is then $\mu|y \sim N(\bar{y}, 1/n)$.
  Contrast with sampling distribution of the sample mean: $\bar{y} \sim N(\mu, 1/n)$.
- The prior $N(\mu_0, n_0^{-1})$ can be interpreted as the posterior obtained from a first sample of the same DGP, of size n_0, with sample mean μ_0, combined with a non-informative prior $\varphi(\mu) \propto 1$. This prior is not a 'proper' (i.e. integrable) density, but the posterior is proper as long as $n_0 \geq 1$.

Density kernels

- When we write $\varphi(\mu|y) \propto \exp[-0.5n_*(\mu - \mu_*)^2]$ the function on the right hand side is called a kernel of the posterior density of μ.
- Similarly $\exp[-0.5n_0(\mu - \mu_0)^2]$ is a kernel of the prior density of μ.
- And $\exp[-0.5n(\mu - \bar{y})^2]$ is a kernel of the data density. Since this density is conditional on μ no factor depending on μ should be forgotten when we drop constants, otherwise we shall make a mistake in computing the posterior! For example, if we assume $y_i|\mu, \sigma^2 \sim I.N(\mu, \sigma^2)$, the relevant kernel is $\sigma^{-n} \exp[-0.5\sigma^{-2} \sum_{i=1}^{n} (y_i - \mu)^2]$.

Less simple example

- Suppose we change the sampling process to $y_i|\mu \sim I.t(\mu, \nu - 2, 1, \nu)$, a Student distribution such that $\mathsf{E}(y_i) = \mu$ and $\mathsf{Var}(y_i) = 1$ (for simplicity, we assume ν known and > 2),
- Or we change the prior density to $\mu \sim t(\mu_0, 3, n_0, 5)$ such that $\mathsf{E}(\mu) = \mu_0$ and $\mathsf{Var}(\mu) = 1/n_0$:
- In both cases, we can write easily the posterior density but it does not belong to a known class and we do not know its moments analytically!
  Note that the ML estimator is not known analytically for the independent t sampling process.

Univariate t (Student) distribution

- A random variable $X \in \mathbb{R}$ has a Student (or t) distribution with parameters $\nu > 0$ (degrees of freedom, $\mu \in R$,) $m > 0$ and $s > 0$, i.e. $X \sim t(\mu, s, m, \nu)$, if its density function is given by

  $$f_t(x|\mu, s, m, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\pi^{\frac{1}{2}}} s^{\frac{1}{2}\nu} m^{\frac{1}{2}} [s + m\,(x - \mu)^2]^{-\frac{1}{2}(\nu+1)}.$$

  Its mean and variance are

  $$\mathsf{E}(X) = \mu \ \text{ if } \nu > 1, \quad \mathsf{Var}(X) = \frac{s}{\nu - 2} m^{-1} \ \text{ if } \nu > 2.$$

- Usual Student quantile tables are for $t(0, \nu, 1, \nu)$ (variance is not 1).

Treatment of parameters

- Classical methods treat the parameters as fixed unknown constants.
- When treating the parameters as random variables, a Bayesian does not believe necessarily that this is reflecting how reality functions, i.e. that θ is randomly drawn from a distribution, and that given this drawn θ, y is drawn from $f(y|\theta)$.
- A more common interpretation is that since θ is an unknown constant, it is possible to make (subjective) probability statements on likely values of θ, resulting in a probability density for θ.

CHAPTER 1

- 1.1 Bayesian inference
- 1.2 Criteria for evaluating statistical procedures
- 1.3 Probability: objective or subjective?
- 1.4 Skills and readings

Criteria for statistical procedures

- Classical statistical procedures are evaluated on their merits in terms of consistency, lack of bias, efficiency... i.e. their properties in hypothetically repeated or large samples.
- Bayesian inference does not care about such properties: it is conditional on the observed sample. All that counts is to use available information ‘coherently’, and not information that might be available but will never be.
- Some Bayesians misleadingly state that Bayesian inference is therefore ‘exact’ in finite samples. This has no meaning since what happens in repeated samples is not relevant for Bayesian inference.

Criteria for statistical procedures

- One may study the sampling properties of any feature of the posterior density, like the expected value, the variance, quantiles...
- Indeed such features are functions of the data and therefore are 'statistics' whose properties depend on the DGP. For example, the posterior mean $\mathsf{E}(\theta|y)$ may be considered as an estimator of θ.
- However, it should not be forgotten that such quantities are ALSO, in principle, depending on the prior! Therefore, comparing the sampling properties of a classical estimator and of the posterior mean should account for the possible prior information embedded in the latter.

Loss function for point estimation

- An estimator (in the classical sense) is a function of the data (a statistic). It depends on the parameter through the DGP.

- Let $l(a, \theta) \geq 0$ be a function measuring the loss of choosing the estimator $a = a(y)$ of the parameter when the value of the parameter is θ. The loss is minimal when $a = \theta$, otherwise it is positive.

  Examples (for scalar θ) are (for $c_1, c_2 > 0$):
  -quadratic loss: $l(a, \theta) = c_1(a - \theta)^2$;
  -piecewise linear loss (asymmetric if $c_1 \neq c_2$):

$$l(a, \theta) = \begin{cases} c_1(a - \theta) & \text{if } a \geq \theta \\ c_2(\theta - a) & \text{if } a < \theta. \end{cases}$$

Bayes point estimators

- A ‘Bayes point estimator’ is a function that minimizes the posterior expected loss:

  $$\theta_* = \arg\min_{a \in \Theta} \mathsf{E}[l(a,\theta)|y],$$

  where $\mathsf{E}[l(a,\theta)|y] = \int l(a,\theta)\varphi(\theta|y)d\theta$.
- Solving this problem for the quadratic loss function yields the posterior mean: $\theta_* = \mathsf{E}(\theta|y)$.
- The solution for the piecewise linear loss function is the $c_1/(c_1 + c_2)$-quantile of the posterior density. In particular, if $c_1 = c_2$, this is the posterior median.

CHAPTER 1

- 1.1 Bayesian inference
- 1.2 Criteria for evaluating statistical procedures
- 1.3 Probability: objective or subjective?
- 1.4 Skills and readings

Probability: objective or subjective?

Kolmogorov’s axiomatic definition of probability is compatible with different interpretations:

- Objective (or empirical): probability as the limit of empirical frequency. It can only apply to events that can be reproduced. There would be no sense to claim that "the return to education coefficient is between 0.04 and 0.08 with probability 0.9".
- Logical: probability is a degree of belief about a proposition, it stems from a logical relation between a proposition and a corpus of knowledge. Consensus regarding this logical link produces a unique system of probabilities.

Probability: objective or subjective?

- Subjective: probability still represents a degree of belief about a proposition but it is no longer based on a universal logical system. It is personal and thus can vary from person to person.
- Bayesian inference, to the extent that it rests on the need to define probabilities on non reproducible events, is hard to reconcile with the empirical interpretation of probability. But there is no need to be a ‘hard core dogmatic subjectivist’ to use Bayesian inference!

CHAPTER 1

- 1.1 Bayesian inference
- 1.2 Criteria for evaluating statistical procedures
- 1.3 Probability: objective or subjective?
- 1.4 Skills and readings

Skills for Bayesian inference

Bayesian inference requires a good knowledge

- of probability distribution theory: it helps to formulate prior distributions, to analyze posterior distributions, and of course to define sensible econometric models (as in classical inference);
  Appendix A of BLR contains a lot of information on this aspect;
- of numerical integration techniques (Ch. 3 of BLR): essential for computing summary features of posterior distributions. Analytical results on posterior densities are limited. This is the counterpart of numerical optimization in classical econometrics.

Main Reference

- Our main reference is "BLR" which stands for Bauwens L., Lubrano, M. and Richard J-F. (1999), *Bayesian Inference for Dynamic Econometric Models*. Oxford University Press.
- For this first chapter, see in particular sections
  -1.4, about interpretations of probability;
  -1.5, about Bayes' Theorem;
  -1.8, about statistical decisional framework and optimality of Bayesian inference rules;
  -1.9, about estimation;
  -1.10, about hypothesis testing;
  -2.3 about kernels.

Other books

- Geweke (2005) Contemporary Bayesian Econometrics and Statistics (Wiley), is a relatively advanced textbook.
- Greenberg (2008), Introduction to Bayesian Econometrics (Cambridge University Press), is a concise introductory textbook.
- Koop (2003), Bayesian Econometrics (Wiley).
- Koop, Poirier and Tobias (2007), Bayesian Econometric Methods (Econometric Exercises 7, Cambridge University Press).

Complementary readings

- Lancaster (2004, Ch. 1) provides a useful synthetic overview of Bayesian inference, including several examples. A topic not covered in BLR is what happens to posterior distributions when the sample size tends to infinity (but see the simple example above, p 12-15).
- Zellner (1971), An Introduction to Bayesian Inference in Econometrics (Wiley) is the first ever published book on Bayesian econometrics. It has a quite interesting chapter (Ch. 2) on the principles and foundations.
- The Oxford Handbook of Bayesian Econometrics (2011) contains chapters on principles, methods, and applications to macroeconomics, microeconomics, marketing, and finance.

COURSE STRUCTURE

- Chapter 1: Concepts (p 2)
- Chapter 2: Numerical Methods
- Chapter 3: Single Equation Regression Analysis (p 95)
- Chapter 4: VAR Models (p 149)

CHAPTER 2

- 2.1 Need for numerical integration
- 2.2 Deterministic integration
- 2.3 Monte Carlo integration

Remark: numerical integration may be useful in classical econometrics also. For some models, the likelihood function or the moment conditions can only expressed as an integral, sometimes of high dimension (like the sample size). This is often the case in models involving latent variables. Examples are the stochastic volatility model, and dynamic discrete choice models.

Bayes theorem using kernels

$$
\begin{aligned}
\varphi(\theta|y) &= \frac{\varphi(\theta)f(y|\theta)}{f(y)} \\
&\propto \varphi(\theta)f(y|\theta) \\
&\propto \kappa(\theta)k(\theta;y) = \kappa(\theta|y).
\end{aligned}
$$

where

- $\kappa(\theta)$ is a kernel of the prior, i.e. $\varphi(\theta) \propto \kappa(\theta)$,
- $k(\theta;y)$ is a kernel of the likelihood function $l(\theta;y) = f(y|\theta)$, i.e. $k(\theta;y)/l(\theta;y)$ must be constant with respect to θ. It is also a good idea to keep in $k(\theta;y)$ all factors depending on y (see next slide);
- $\kappa(\theta|y)$ is a kernel of the posterior.

Need for numerical integration 1

- Case 1: when we don't know the density corresponding to $\kappa(\theta|y)$, we don't know analytically the constant $K(y)$ such that $\varphi(\theta|y) = \kappa(\theta|y)/K(y)$ is a properly normalized density, i.e. such that $\int \varphi(\theta|y)d\theta = 1$.

  Obviously, $K(y) = \int \kappa(\theta|y)d\theta = \int \kappa(\theta)k(\theta;y)d\theta$.
  Note that $f(y) \propto K(y)$ if we include in $k(\theta;y)$ all factors depending on y.

  Still in case 1, we don't know analytically posterior moments. The posterior expectation of an integrable function $g(\theta)$ is defined as:

$$\mathsf{E}[g(\theta)|y] = \int g(\theta)\varphi(\theta|y)d\theta$$

Need for numerical integration

$$\mathsf{E}[g(\theta)|y] = \int g(\theta)\varphi(\theta|y)d\theta$$

and must be computed as:

$$\mathsf{E}[g(\theta)|y)] = \frac{\int g(\theta)\kappa(\theta)k(\theta;y)d\theta}{\int \kappa(\theta)k(\theta;y)d\theta}.$$

Notice that the denominator is a particular case of the numerator, when $g(\theta) = 1$. Therefore we need to compute integrals of the type

$$\int g(\theta)\kappa(\theta)k(\theta;y)d\theta$$

for several functions $g(.)$, depending on what posterior results we want to report.

Functions and their expectations

$g(\theta)$	$\mathsf{E}[g(\theta)\vert y]$
θ	posterior mean
$\theta\theta'$	matrix of uncentered second order moments
$1_{\theta\in A}$	posterior probability of event A
$f(y_h^*\vert\theta,y)$	predictive density of y_h at value y_h^*
$\mathsf{E}(y_h\vert\theta,y)$	predictive mean of y_h

The posterior variance-covariance matrix is computed as $\mathsf{E}(\theta\theta'|y) - \mathsf{E}(\theta|y)\mathsf{E}(\theta'|y)$.

Note that $\mathsf{E}(y_h|y) = \int \mathsf{E}(y_h|\theta,y)\varphi(\theta|y)d\theta$ (law of iterated expectations). Likewise, $f(y_h^*|y) = \int f(y_h^*|\theta,y)\varphi(\theta|y)d\theta$.

Marginal densities

By defining the set A appropriately, and computing $\mathsf{E}(1_{\theta\in A}) = \mathsf{Pr}(\theta \in A)$, one can approximate the ordinates of the marginal density of any element of θ. Suppose θ is scalar and $A = (a, b)$, with $b - a$ small. Then

$$\int_a^b \varphi(\theta|y)d\theta = \mathsf{Pr}(\theta \in A) \simeq \varphi\Big(\frac{a+b}{2}|y\Big)(b-a)$$

This should be done for a fine grid of abscissae of θ where the marginal posterior has its mass.
The procedure can be extended to compute the marginal posterior of two parameters by defining A as a small rectangle.

Need for numerical integration 2

- Case 2: even if we know the normalizing constant, moments, and marginal densities of $\varphi(\theta|y)$, we may not know analytically the posterior density or the moments of some functions g.

  Example 1: suppose $\theta|y \sim N_2(\theta_*, V_*)$ but we want to compute the posterior density of θ_1/θ_2. Note however that this particular function does not have a finite mean, but it has a density which can be computed from the joint density of θ.

  Example 2: the roots of the determinantal equation of a VAR model are highly non-linear functions of the parameters.

Chapter 2

- 2.1 Need for numerical integration
- 2.2 Deterministic integration
- 2.3 Monte Carlo integration

Deterministic integration

- Useful and easy for computing an integral of dimension 1 or 2. Let θ be a scalar.
- Let $I = \int_0^1 h(\theta) d\theta$ be the integral to compute. Typically, $h(\theta) = g(\theta)\kappa(\theta|y)$.
- The limits of integration are in general not 0 and 1 but it is always possible to make a change of variable to express the problem as above. For example: $\int_a^b h(\theta) d\theta = (b-a) \int_0^1 h[(b-a)\tau + a] d\tau$.
- When a is $-\infty$ and/or b is ∞, another transformation must be used. For example $\tau = 1/(1+\exp(-\theta)) \in (0,1)$ if $\theta \in (-\infty, \infty)$.

Trapezoidal rule

- Approximate $h(\theta)$ by a linear function through the endpoints and deliver the area of the trapezium as the integral: $I = \int_0^1 h(\theta)\, d\theta \simeq \dfrac{h(0)+h(1)}{2}$.
- In practice: split $(0,1)$ into $2n$ intervals of equal length based on the points $\theta_0 (= 0), \theta_1, \ldots, \theta_{2n} (= 1)$, apply the rule to each interval and add the pieces:

$$I \simeq \frac{1}{4n}[h(\theta_0) + 2h(\theta_1) + 2h(\theta_2) + \cdots + 2h(\theta_{2n-1}) + h(\theta_{2n})]$$

- The error is proportional to $(2n+1)^{-2}$. Use $n = 100$ at least.

Simpson's rule

- Approximate $h(\theta)$ by a quadratic function through $h(0), h(0.5), h(1)$ and deliver the area under the quadratic function as the integral: $I \simeq [h(0) + 4h(0.5) + h(1)]/6$.
- Using $2n + 1$ equally spaced points:

$$\begin{aligned} I \;\simeq\; & \frac{1}{6n}[h(\theta_0) + 4h(\theta_1) + 2h(\theta_2) + 4h(\theta_3) + 2h(\theta_4) \\ & + \cdots + 2h(\theta_{2n-2}) + 4h(\theta_{2n-1}) + h(\theta_{2n})] \end{aligned}$$

- The error is proportional to $(2n + 1)^{-4}$. Use $n = 8$ at least. For given n, Simpson's rule should be more precise than trapezoidal rule. Use Simpson's rule if function evaluation is costly.

Chapter 2, Section 2.3

- 2.1 Need for numerical integration
- 2.2 Deterministic integration
- 2.3 Monte Carlo integration
  - 2.3.1 Definition
  - 2.3.2 Independent sampling
  - 2.3.3 Dependent sampling

Monte Carlo integration

- Useful, but not always easy, for computing an integral of dimension ≥ 3. Can also be used for smaller dimensions but deterministic rules are more efficient.
- Deterministic rules cannot be used for a large dimension: for a dimension of k and with a grid of G points for each coordinate, we need G^k function evaluations: e.g. $20^9 \times 10^{-6}$ implies 6 days of computing; for $k = 10$, it becomes 118 days! And the programming would not be easy.
- Monte Carlo integration uses much less points but chooses them where they are most useful, i.e. where the integrand varies a lot.

Principle of Monte Carlo integration

- The general principle of a Monte Carlo method is:
  (i) to express the solution of a problem as a parameter of a hypothetical population,
  (ii) to use random numbers to build a (dependent or independent) sample of the population, and
  (iii) to estimate the parameter of the population using the generated sample.
- See Appendix B of BLR for random number generation from many probability distributions.
- Ironically, estimation is done in the classical (i.e. non-Bayesian) sense, and justified by the consistency of the estimator for the parameter value...

Simple example

- Let $\theta \sim N_2(\mu, \Sigma)$ be a posterior density.
- We want to compute $\text{Pr}(\theta \in A)$, and the median and density of θ_1/θ_2.
- Let $\{\theta^{(i)}\}_{i=1}^n$ be an I.I.D. simulated sample from the posterior. See p 319 of BLR for an algorithm to generate from a multivariate normal, and p 317 for a $N(0, 1)$.
- We estimate $\text{Pr}(\theta \in A)$ by the proportion of simulated $\theta^{(i)} \in A$.
- With the sampled value, we can approximate the density of θ_1/θ_2 by a kernel method. The median is estimated by the sample median.

Chapter 2, Section 2.3

- 2.3 Monte Carlo integration
  - 2.3.1 Definition
  - 2.3.2 Independent sampling
    Direct sampling
    Importance Sampling
    Rejection sampling
  - 2.3.3 Dependent sampling

Direct sampling

- Hypothetical population: $\theta \sim \varphi(\theta)$
- Let $\{\theta^{(i)}\}_{i=1}^{n}$ be an I.I.D. simulated sample of that population.
- $g_D = \sum_{i=1}^{n} g(\theta^{(i)})/n$ is unbiased and consistent for $\mu_g \equiv \mathrm{E}[g(\theta)]$.
- $n^{1/2}(g_D - \mu_g) \overset{a}{\sim} N(0, \sigma_g^2)$, where $\sigma_g^2 < \infty$ is the population variance of $g(\theta)$. Hence, with probability $1-\alpha$, $|\frac{g_D}{\mu_g} - 1| < z_\alpha \frac{\sigma_g}{\mu_g} \frac{1}{\sqrt{n}}$ (z_α is the quantile such that $\Pr(|Z| < z_\alpha) = 1 - \alpha$, Z being standard normal).
- $g_D \pm z_\alpha(s_g/\sqrt{n})$ is an estimated confidence interval of level $1-\alpha$, where s_g is the sample standard deviation of $g(\theta)$.

Importance sampling (IS)

- With a posterior in an unknown class, direct sampling is not applicable because we cannot sample directly from the posterior.
- Importance sampling changes the hypothetical population $\varphi(\theta)$ by another population $\iota(\theta)$ (called importance function or density) wherefrom an I.I.D. sample can be generated and computes the expectation as a parameter of the new population. We wish to compute

$$\mu_g = \mathsf{E}_{\varphi}[g(\theta|y)].$$

  Note that:

Fundamental IS equality

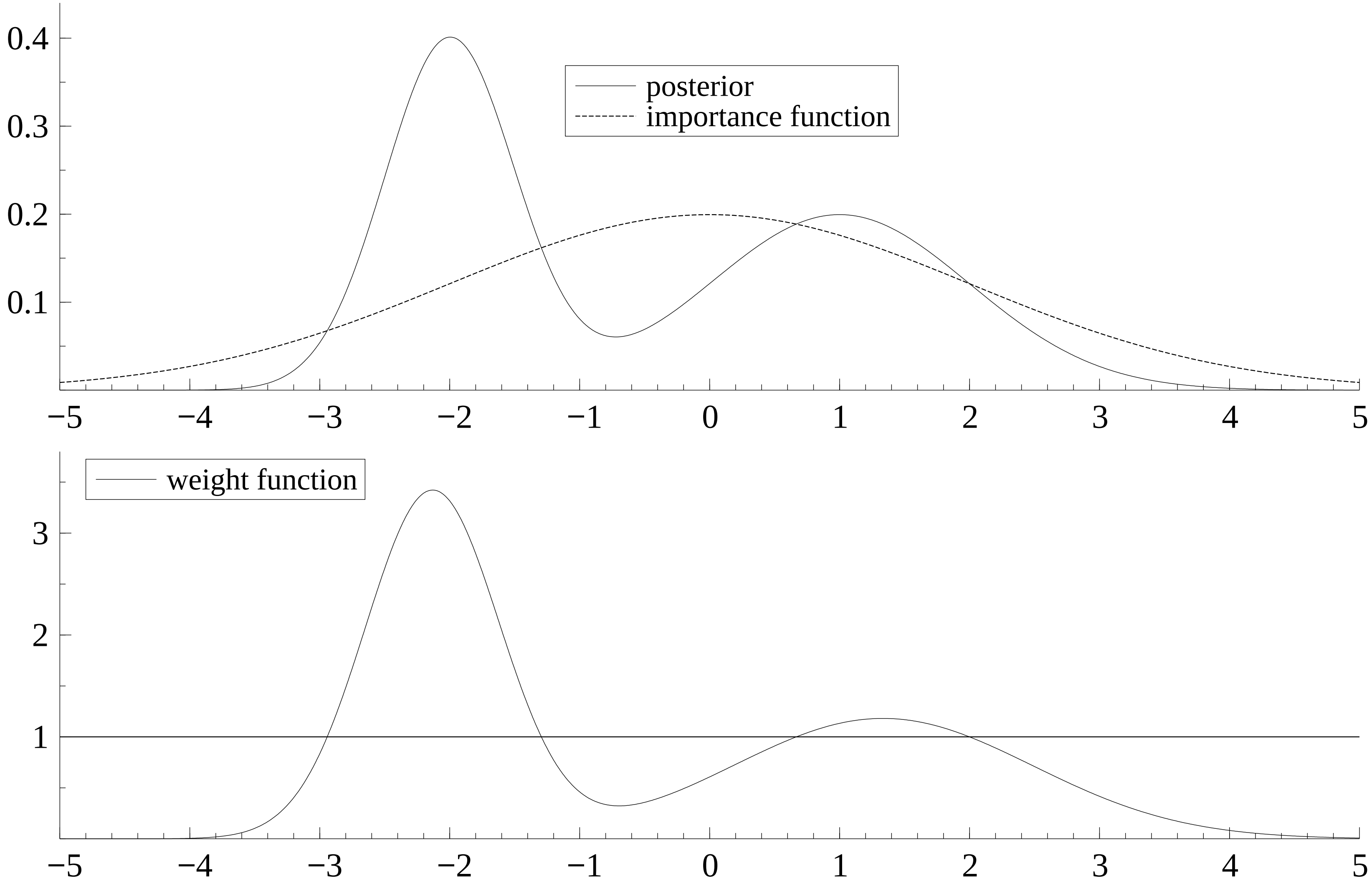
$$\mu_g = \mathsf{E}_{\varphi}[g(\theta|y)] = \int g(\theta)\varphi(\theta|y)d\theta = \frac{\int g(\theta)\kappa(\theta|y)d\theta}{\int \kappa(\theta|y)d\theta}$$

$$\begin{aligned}\int g(\theta)\kappa(\theta|y)d\theta &= \int g(\theta)\frac{\kappa(\theta|y)}{\iota(\theta)}\iota(\theta)d\theta \\ &= \mathsf{E}_{\iota}[g(\theta)\frac{\kappa(\theta|y)}{\iota(\theta)}] = \mathsf{E}_{\iota}[g(\theta)w(\theta)]\end{aligned}$$

$$\int \kappa(\theta|y) = \mathsf{E}_{\iota}[w(\theta)]$$

$$\Rightarrow \mu_g = \frac{\mathsf{E}_{\iota}[g(\theta)w(\theta)]}{\mathsf{E}_{\iota}[w(\theta)]}$$

IS graph



0.4
0.3
0.2
0.1
posterior
importance function
−5
−4
−3
−2
−1
0
1
2
3
4
5
weight function
3
2
1
−5
−4
−3
−2
−1
0
1
2
3
4
5

Estimation

- Let $\{\theta^{(i)}\}_{i=1}^{n}$ be an I.I.D. sample of $\iota(\theta)$.
- μ_g is estimated by

$$g_I = \frac{\frac{1}{n}\sum_{i=1}^{n} g(\theta^{(i)})\, w(\theta^{(i)})}{\frac{1}{n}\sum_{i=1}^{n} w(\theta^{(i)})} = \sum_{i=1}^{n} \omega^{(i)}\, g(\theta^{(i)})$$

  where $\omega^{(i)} = w(\theta^{(i)})/\sum_{j=1}^{n} w(\theta^{(j)})$.

- Expectations are computed as weighted averages of the incorrect sample but the sampled values are weighted in order to get a consistent estimator.
- $w(\theta) = \frac{\kappa(\theta|y)}{\iota(\theta)}$ is called the weight function.

Properties of g_I

- Note that g_I is a ratio of unbiased and consistent estimators (for the numerator and denominator) but is not unbiased for μ_g (because of the ratio). However it is consistent.
- Under regularity conditions, $n^{1/2}(g_I - \mu_g) \overset{a}{\sim} N(0, \tau_g^2)$, where $\tau_g^2 < \infty$ is the asymptotic variance of g_I.
- For scalar θ and $g(\theta) = \theta$, BLR (p 77-78) show how to estimate consistently τ_g^2.
- Using this, the probabilistic bound on the error $|\frac{g_I}{\mu_g} - 1|$, and a confidence interval for μ_g can be estimated as in the case of direct sampling.

How to choose $\iota(\theta)$?

- For $\bar{w}_n = \frac{1}{n}\sum_{i=1}^{n} w(\theta^{(i)})$ to be consistent for $\mathsf{E}_{\iota}[w(\theta)]$, the latter should be finite. And the smaller the variance of $w(\theta)$, the more precise $\bar{w}_n$.
- Obviously, if $\iota(\theta) \propto \kappa(\theta)$, $w(\theta)$ is a finite constant hence has zero variance. In this case we would be sampling from $\varphi(\theta)$ and back to the case of direct sampling.
- Fundamental guideline: choose $\iota(\theta)$ as close as possible to $\kappa(\theta)$ and such that $w(\theta) < \infty$.
- If $\iota(\theta)$ provides a good approximation to $\int \kappa(\theta|y)d\theta$ $= \mathsf{E}_{\iota}[w(\theta)]$, it will also provide a good approximation to $\mathsf{E}_{\iota}[g(\theta)w(\theta)]$ and even a better one to μ_g (see BLR, p 79).

Useful information

One should use as much as possible the information one has about the posterior density in order to build an importance function (I.F.). The type of information that may be available is theoretical (Ti) or empirical (Ei):

T1: conditions on the existence of moments;

T2: existence of one or several modes;

T3: characterization of some conditional densities of φ including their moments;

E1: mode and Hessian of $\log(\varphi)$ evaluated at the posterior mode (by numerical optimization);

E2: a first approximation of the moments of φ (by Laplace's method, or a normal approximation, or a first round of importance sampling).

General hints on ι

Select ι so that it matches the location, covariance structure, and tail behaviour of φ.

1. Location: ι should have the same mode(s) as φ (using T2, and E1 or E2).
2. An approximation to the covariance matrix of φ is given by minus the Hessian inverse of $\log(\varphi)$ (see E1) or by E2. It is useful to inflate the approximate covariance matrix a little.
3. Tails: ι should have moments of order not higher than φ (using T1). If ι has thicker tails than φ, this avoids extreme values of $w(\theta)$ in the tails.

   2. and 3. help to prevent the explosion of the weight function in the tails of ι.

Method 1

Normal or Student approximation around the mode:

- A 2nd-order Taylor expansion of $\log(\varphi)$around its mode θ_* is equal to constant $- 0.5(\theta - \theta_*)'H(\theta - \theta_*)$, where $H = -\partial^2 \log(\varphi)/\partial\theta\partial\theta'|_{\theta_*}$. Hence $\theta \sim N(\theta_*, cH^{-1})$, where $c \geq 1$ is a tuning constant is a possible importance density.
- To get thicker tails, use $\theta \sim t(\theta_*, 1, H/c, \nu)$ with $\nu \leq$ order of existence of posterior moments (if T1).
- Not appropriate for very skewed or multimodal posterior! Bauwens and Laurent (JBES, 2005) show how to make multivariate normal or Student densities skewed.

Method 2

Importance function incorporating exact conditional densities:

- If $\theta = (\alpha \, \beta)$ and $\varphi(\beta|\alpha)$ can be simulated directly, it should be incorporated in $\iota(\theta)$, i.e.

$$\iota(\theta) = \varphi(\beta|\alpha) \, \iota_m(\alpha),$$

  where $\iota_m(.)$ is the marginal I.F. of α, an approximation of $\varphi(\alpha)$ (obtained by another method).

- Then, the weight function depends only on α.
- Simulation of ι is done sequentially: $\alpha^{(i)}$ is drawn from $\iota_m(.)$, then β from $\varphi(\beta|\alpha^{(i)})$. Repeat n times.

Rao-Blackwellisation

- Instead of using

$$\beta_I = \frac{\sum_{i=1}^n \beta^{(i)} w(\alpha^{(i)})}{\sum_{i=1}^n w(\alpha^{(i)})}$$

to estimate $\mathsf{E}(\beta|y)$, if $\mathsf{E}(\beta|\alpha, y)$ is known analytically and easy to compute, one can use the estimator

$$\beta_{Ic} = \frac{\sum_{i=1}^n \mathsf{E}(\beta|\alpha^{(i)}, y)\, w(\alpha^{(i)})}{\sum_{i=1}^n w(\alpha^{(i)})}$$

of $\mathsf{E}_{\alpha|y}[\mathsf{E}(\beta|\alpha, y)] = \mathsf{E}(\beta|y)$. This is more efficient because $\mathsf{Var}_{\alpha|y}[\mathsf{E}(\beta|\alpha, y)] \leq \mathsf{Var}(\beta|y)$. Indeed,

Rao-Blackwellisation

- $\mathsf{Var}(\beta|y) = \mathsf{E}_{\alpha|y}[\mathsf{Var}(\beta|\alpha, y)] + \mathsf{Var}_{\alpha|y}[\mathsf{E}(\beta|\alpha, y)]$.
- The latter relation can also be used to estimate $\mathsf{Var}(\beta|y)$ by

$$\frac{\sum_{i=1}^{n} \mathsf{Var}(\beta|\alpha^{(i)})\, w(\alpha^{(i)})}{\sum_{i=1}^{n} w(\alpha^{(i)})} + \frac{\sum_{i=1}^{n} \mathsf{E}(\beta|\alpha^{(i)})\, \mathsf{E}(\beta'|\alpha^{(i)})\, w(\alpha^{(i)})}{\sum_{i=1}^{n} w(\alpha^{(i)})} - \beta_{Ic}\beta'_{Ic}$$

- For the marginal density at the value β:

$\varphi(\beta) = \mathsf{E}_{\alpha}[\varphi(\beta|\alpha)] \simeq \sum_{i=1}^{n} \varphi(\beta|\alpha^{(i)})\, w(\alpha^{(i)}) / \sum_{i=1}^{n} w(\alpha^{(i)})$. Repeat this for a grid of β to plot the density.

Method 3

Optimal choice of the parameters of the I.F.:

- First choose a parametric family for the I.F. (normal, Student, skew-Student...). Denote it by $\iota(\theta|\lambda)$ where λ is the parameter vector of the I.F.
- Then minimize the Monte Carlo variance of the quantity to be estimated. Choosing this to be the integral of the posterior kernel (c_κ), one has to find λ that minimizes Var$_\iota(\kappa/\iota)$ or equivalently $\int \frac{[\kappa(\theta)]^2}{\iota(\theta|\lambda)} d\theta$, but the integral is usually not known. A numerical approximation may be available. This method is used in efficient importance sampling.

Other methods

- In method 2, $\iota_m(\alpha)$ could be $\varphi(\alpha|\hat{\beta})$ where $\hat{\beta}$ is the posterior mode of β, if this is a known density that can be easily simulated. Beware of too small variances due to conditioning (Rao-Blackwell!).
- Sequential updating: start with an I.F. obtained by method 1 or 2, estimate posterior moments with it, then define a new I.F. that uses these first estimates of the posterior moments. This helps to improve the covariance matrix and to discover important skewness directions in the posterior.
- A transformation of θ may induce more symmetry (e.g. with skewness to the right, the log induces more symmetry).

Note on the weight function

- In practice, use a kernel $i(\theta)$ of the I.F., such that $\iota(\theta) = i(\theta)/c_i \propto i(\theta)$ where $c_i = \int i(\theta) d\theta$ is analytically known.
- Define $w(\theta) = \kappa(\theta|y)/i(\theta)$, rather than $\kappa(\theta|y)/\iota(\theta)$.
- Then $\bar{w}_n = \frac{1}{n}\sum_{i=1}^{n} w(\theta^{(i)})$ estimates c_κ / c_i, where $c_\kappa = \int \kappa(\theta|y) d\theta$, since

  $$1 = \int \frac{c_\kappa^{-1}\kappa(\theta|y)}{c_i^{-1} i(\theta)} \iota(\theta) d\theta = \frac{c_\kappa^{-1}}{c_i^{-1}} \mathsf{E}_\iota\left[\frac{\kappa(\theta|y)}{i(\theta)}\right] \simeq \frac{c_\kappa^{-1}}{c_i^{-1}} \bar{w}_n$$

  $\Rightarrow \bar{w}_n c_i$ is an operational estimator of c_κ.
- For computing $\mu_g = \frac{\mathsf{E}_\iota[g(\theta)w(\theta)]}{\mathsf{E}_\iota[w(\theta)]}$, using $i(\theta)$ rather than $\iota(\theta)$ in w does not make any difference.

Practical convergence hints

- Always check the weights: plot their histogram, and be sure that there is not just a single draw that determines the results. This happens if there is a draw with a relative weight close to 1.
- Check the coefficient of variation of the weights. It should stabilize when n is increased. If it explodes, the I.F. is not good enough.
- Always estimate the probabilistic error bound of the estimates of the integral of the posterior kernel (c_κ) and of the posterior expectations. They should not exceed 5 to 10 per cent. With a stable coefficient of variation of the weights, this can always be attained by increasing n.

Rejection sampling (RS)

- Instead of weigthing the draws of an I.F., RS checks if a draw generated from it is acceptable as a draw of the posterior.
- A candidate draw θ is accepted if

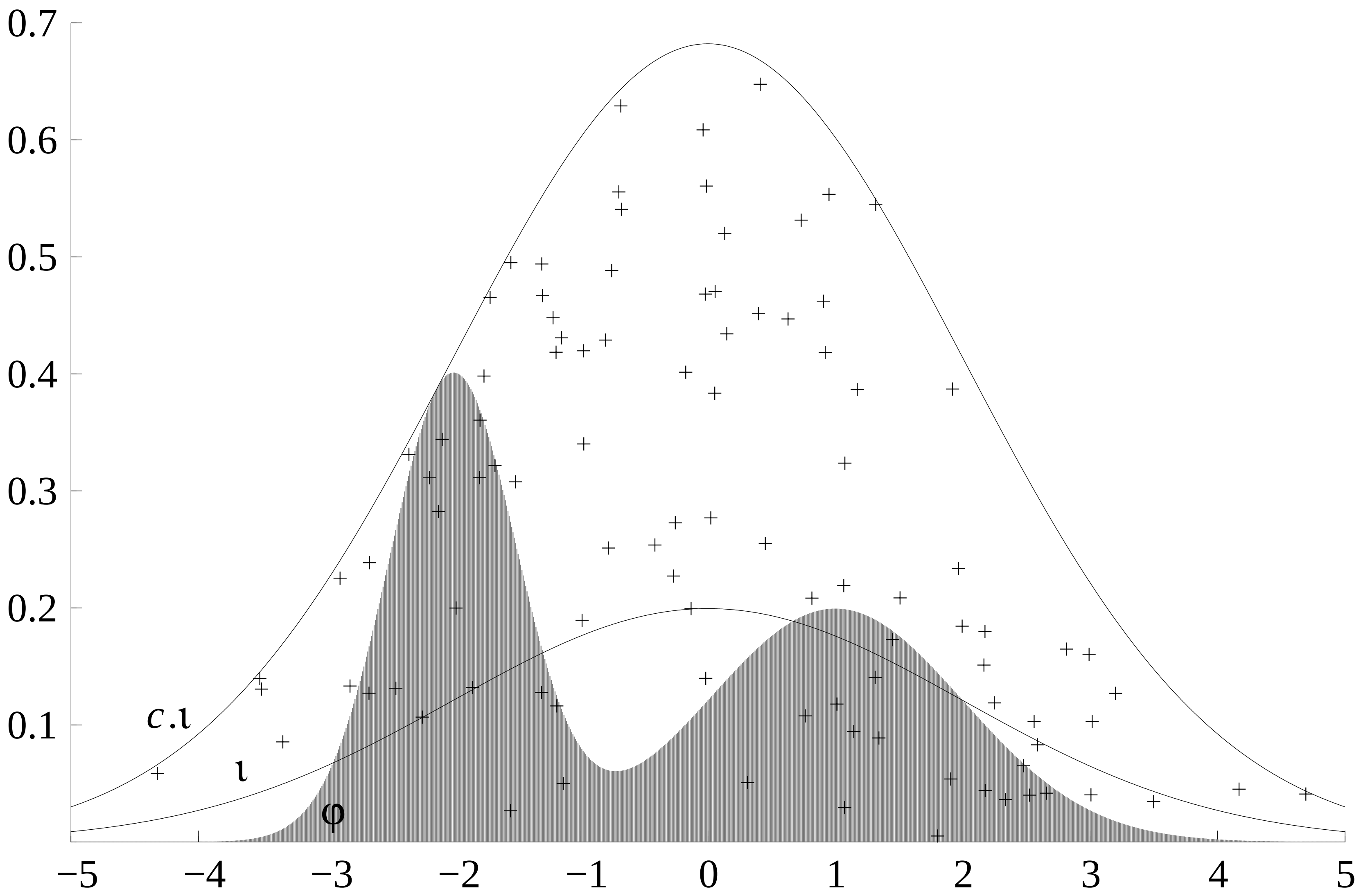
$$\kappa(\theta) > c.i(\theta)u,$$

where $u \in (0,1)$ is a uniform random number, and

$c \geq \sup_{\theta} \frac{\kappa(\theta)}{i(\theta)} \in [1,\infty).$

The purpose of multiplying i by c is to envelope κ. The volume between κ and the envelope is the rejection region and should be as small as possible (hence set c = the sup).

RS graph



RS: properties

- Accepted draws constitute an IID sample of the posterior (see next slide) and can be used as in direct sampling to estimate posterior moments.
- The percentage of accepted draws is an estimator of $c_\kappa/(c.c_i)$. This percentage can be very low, and the method can be very inefficient in the sense that the average computing time to get one accepted draw is very large.
- For a given ι and g, the importance sampling estimator of μ_g has a smaller Monte Carlo variance than the rejection sampling estimator.
- Mainly used for simulating some univariate distributions by building a tight envelope.

RS: proof

The density of an accepted θ^* is the density $\iota(.)$ conditional on the acceptation of the generated point:

$$\iota(\theta|u \leq \kappa(\theta^*)/c.i(\theta^*))$$

$$= \frac{\text{Pr}[u \leq \kappa(\theta^*)/c.i(\theta^*)|\theta^* = \theta]\iota(\theta)}{\text{Pr}[u \leq \kappa(\theta^*)/c.i(\theta^*)]} \text{ (by Bayes theorem)}$$

$$= \frac{[\kappa(\theta)/c.i(\theta)]\iota(\theta)}{\int \text{Pr}[u \leq \kappa(\theta^*)/c.i(\theta^*)|\theta^* = \theta]\iota(\theta)d\theta} \text{ (since } u \sim U(0,1))$$

$$= \frac{[\kappa(\theta)/c.i(\theta)]\iota(\theta)}{\int [\kappa(\theta)/c.i(\theta)]\iota(\theta)d\theta} = \frac{\kappa(\theta)/(c.c_i)}{\int [\kappa(\theta)/(c.c_i)]d\theta} = \varphi(\theta)$$

(since $\iota(\theta)/i(\theta) = c_i^{-1}$ and $c.c_i$ is a constant).

RS: special simple case

- When the prior implies inequality constraints, say $\varphi(\theta) \propto \kappa(\theta)\mathbb{1}_{\{\theta\in\Theta_r\}}$ and $\Theta_r \in \Theta$ where Θ is the unrestricted parameter space, and the posterior without constraints is easy to simulate (directly or otherwise), the unconstrained posterior can be used as the simulator ι, and one has simply to reject the draws that do not lie in Θ_r.
- Example: $y_t = x_t'\beta + \epsilon_t,\ \epsilon_t \sim I.N(0,\sigma^2),\ t = 1,\ldots,T.$ Prior: $\varphi(\beta,\sigma^2) \propto \mathbb{1}_{\{\beta_1>0\}}1/\sigma^2$.
  Posterior: $\beta|d \sim t(\hat{\beta}, y'My, X'X, T-k)\mathbb{1}_{\{\beta_1>0\}} \Rightarrow$ generate from the t and keep only draws such that $\beta_1^{(i)} > 0$.
  Beware Pr$(\beta_1 > 0|d)$ is small!

Chapter 2, Section 2.3

- 2.3 Monte Carlo integration
  - 2.3.1 Definition
  - 2.3.2 Independent sampling Direct sampling
    Importance Sampling
    Rejection sampling
  - 2.3.3 Dependent sampling
    Independent Metropolis-Hastings sampling
    Random walk MH sampling
    Gibbs sampling
    Estimation of posterior moments and diagnostics

MCMC

- Markov chain Monte Carlo (MCMC) methods produce dependent samples of the posterior, rather than independent samples produced by direct, importance or rejection sampling. However, they can deal with case 1 (unknown type of posterior) and are more easy to apply than importance or rejection sampling in many cases.
- Dependence in the sample does not impede consistent estimation of posterior features provided the generated sequence is ergodic.
  However dependence renders statistical inference more difficult, e.g. to estimate the variance of the mean of a dependent sample is more difficult than for an independent sample.

MCMC

- Convergence is not granted and must be checked. The question of convergence is whether the sequence of generated draws can be considered as draws of the posterior, i.e. does the distribution of the draws generated by a MCMC sampler converge to the posterior?
- Moreover, there is no ‘general’ asymptotic normality theorem for ergodic sequences. Appropriate conditions are ‘case-dependent’.
- Gibbs sampling and Metropolis-Hastings (MH) sampling are the two most important classes of MCMC algorithms. They can be combined.

Independent MH sampling

- As IS and RS, MH uses an auxiliary density, called the candidate density, to generate θ. It uses an acceptance/rejection mechanism to decide if a draw can/cannot be accepted as a draw of the posterior. Let $\theta^{(i)}$ be the last accepted draw.
- The next draw $\theta^{(i+1)}$ is generated as follows:

  1) generate $\theta^{*} \sim \iota(\theta)$

  2) compute $p = \min\left[\frac{\varphi(\theta^{*})}{\varphi(\theta^{(i)})}\frac{\iota(\theta^{(i)})}{\iota(\theta^{*})}, 1\right] = \min\left[\frac{w(\theta^{*})}{w(\theta^{(i)})}, 1\right]$

  3) take $\theta^{(i+1)} = \begin{cases} \theta^{*} & \text{with probability } p \\ \theta^{(i)} & \text{with probability } 1-p. \end{cases}$

Independent MH sampling

- If $\varphi = \iota$, this is direct sampling.
- To compute p, there is no need to know the integrals of the posterior and importance kernels.
- θ^*, the candidate draw, is accepted surely if it has more weight than the previous draw. Otherwise, it is accepted with probability $w(\theta^*)/w(\theta^{(i)})$.
- Some draws may be repeated several times, i.e. $\theta^{(i+1)} = \theta^{(i)} = \theta^{(i-1)} = ...$ may occur. This creates dependence in the sample.
- This algorithm is called the *independent* MH, because the candidate density does not depend on the previous accepted draw.

Random walk MH sampling

- The RW MH sampler generates a candidate draw θ^* by a move around $\theta^{(i)}$: $\theta^* = \theta^{(i)} + \epsilon$ where ϵ is a draw of a density $\iota(\epsilon)$ centered on 0 and with an appropriate covariance structure. Hence $\theta^* \sim \iota(\theta|\theta^{(i)})$. Often, the density of ϵ is chosen as multivariate normal with covariance Σ, so that $\theta^* \sim N(\theta^{(i)}, \Sigma)$. The choice of Σ is crucial.
- The next draw $\theta^{(i+1)}$ is obtained as follows:

  1) generate $\theta^* \sim \iota(\theta|\theta^{(i)})$

  2) compute $p = \min\left[\dfrac{\varphi(\theta^*)}{\varphi(\theta^{(i)})}\dfrac{\iota(\theta^{(i)}|\theta^*)}{\iota(\theta^*|\theta^{(i)})}, 1\right]$

  3) take $\theta^{(i+1)} = \begin{cases} \theta^* & \text{with probability } p \\ \theta^{(i)} & \text{with probability } 1-p. \end{cases}$

Convergence and ergodicity

- Sufficient conditions for convergence and ergodicity of a MH sampler when the candidate $\iota(\theta|\theta^{(p)})$ depends on the previous draw $\theta^{(p)}$, are that i) the candidate is $> 0 \ \forall \theta$ and $\theta^{(p)} \in \Theta$, and
  ii) $\varphi(\theta|y) > 0 \ \forall \theta$ in the parameter space Θ.
- For an independent MH sampler, it is sufficient that $0 < w(\theta) < \infty \ \forall \theta \in \Theta$.
  In this case, one can even invoke an asymptotic normality theorem for the simple average of an (integrable) function of the draws (because the draws are uniformly ergodic).

Simple example

- Suppose $\varphi(\theta)$ is a $N(0,1)$ and $\iota(\theta)$ is a $N(0,1/c)$.
- $\Rightarrow w(\theta) = c^{-1/2} \exp[0.5(c-1)\theta^2]$ and $w(\theta)/w(\theta') = \exp[0.5(c-1)(\theta^2 - \theta'^2)]$
- Suppose that $c = 3$, i.e. the candidate is more concentrated than the target.
- A move from $\theta' = 2$ to $\theta = 0$ is very unlikely since $p = w(2)/w(0) = \exp(-4) = 0.018$: the value 2 must be 'oversampled'.
- A move from $\theta' = 0$ to $\theta = 2$ is surely accepted since $w(0)/w(2) = \exp(4) > 1$ so that $p = 1$.

Gibbs sampling

- Let $\varphi(\theta)$ denote the posterior. Suppose that one can partition θ into two ‘blocks’ as (θ_1, θ_2) such that the ‘full conditional densities’ $\varphi(\theta_1|\theta_2)$ and $\varphi(\theta_2|\theta_1)$ can be directly simulated (e.g. they are normal densities, although the joint density is not normal).
- The Gibbs sampler generates a sequence of dependent draws of the joint posterior as follows: given $\theta^{(i-1)}$, the next point $(\theta_1^{(i)}, \theta_2^{(i)})$ is generated by the following cycle:

$$\begin{aligned} \theta_1^{(i)} &\sim \varphi(\theta_1|\theta_2^{(i-1)}), \\ \theta_2^{(i)} &\sim \varphi(\theta_2|\theta_1^{(i)}). \end{aligned}$$

Gibbs sampling

- An initial value $\theta_2^{(0)}$ (in the support of the posterior) must be provided, which has no influence after a sufficiently large number of cycles has been performed (if the Markov chain sequence is ergodic).
  $\Rightarrow$ Always discard a ‘warm-up’ (or ‘burn-in’) sequence of initial draws to get rid of the influence of the initial value.
- Pick the initial value as a central value (from the posterior mode for example) rather than an unusual value far in the tails.

Convergence and ergodicity

- Loosely stated, a sufficient condition for convergence and ergodicity is that there is a positive probability to move from any point to any other point (the full conditional densities must be strictly positive for all values of the conditioning variables).
- Another sufficient condition excludes ‘pathological’ cases where the sampler gets trapped in a point or a subset of the parameter space. For example, with a disconnected space, the danger is to never visit a part of the space.

Difficulty of Gibbs sampling

- A lot of dependence in the chain, implying that a large number of draws are needed to explore the posterior. A great danger is to remain stuck for long in a part of the parameter space.
- Example: let $(\theta_1, \theta_2)' \sim N_2(0, R)$ where R is a correlation matrix with ρ the off-diagonal element.
  $\Rightarrow \theta_i|\theta_j \sim N(\rho\theta_j, 1 - \rho^2)$.
  If $|\rho| = 1$, the chain stays always at the starting point. If $|\rho|$ is close to 1, it moves quite slowly in the space of θ.
- Highly correlated elements of θ should be in the same block, if possible!
  NB: one block corresponds to direct sampling.

More blocks

- With m blocks, i.e. $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$, a draw $\theta^{(i)}$ is generated using the previous draw $\theta^{(i-1)}$ by the following cycle:

$$
\begin{aligned}
\theta_1^{(i)} &\sim \varphi(\theta_1 | \theta_2^{(i-1)}, \ldots, \theta_m^{(i-1)}), \\
\theta_2^{(i)} &\sim \varphi(\theta_2 | \theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_m^{(i-1)}), \\
&\vdots \\
\theta_j^{(i)} &\sim \varphi(\theta_j | \theta_1^{(i)}, \ldots, \theta_{j-1}^{(i)}, \theta_{j+1}^{(i-1)}, \ldots, \theta_m^{(i-1)}), \\
&\vdots \\
\theta_m^{(i)} &\sim \varphi(\theta_m | \theta_1^{(i)}, \ldots, \theta_{m-1}^{(i)}).
\end{aligned}
$$

Combining Gibbs and MH

- Assume 2 blocks, such that $\theta_1|\theta_2$ can be simulated directly but $\theta_2|\theta_1$ not.
- In the Gibbs sampling algorithm, the step of drawing $\theta_2|\theta_1$ can be implemented by an indirect method, like rejection or MH sampling. The candidate density should be redefined at each iteration of the Gibbs to take account of the last sampled value of θ_1.
- Example: nonlinear regression:
  $y_t = f(x_t'\beta) + \epsilon_t,\ \epsilon_t \sim I.N(0,\sigma^2),\ t = 1,\ldots,T.$
  Prior: $\varphi(\beta,\sigma^2) \propto \varphi(\beta)1/\sigma^2$.
  Posterior: $\beta|\sigma^2, d$ in unknown class,
  $\sigma^2|\beta, d \sim IG_2(\sum_{t=1}^T \epsilon_t^2, T)$.

Estimation of posterior moments

- $g_{mcmc} = \sum_{i=1}^{n} g(\theta^{(i)})/n \xrightarrow{p} \mu_g \equiv \mathsf{E}[g(\theta)]$ as in direct sampling.
- In Gibbs sampling, 'Rao-Blackwellisation' can be used, e.g. with 2 blocks $\sum_{i=1}^{n} \mathsf{E}(\theta_1|\theta_2^{(i)})/n \xrightarrow{p} \mathsf{E}(\theta_1)$. However, there is no guarantee that this estimator has smaller variance than $\sum_{i=1}^{n} \theta_1^{(i)}/n$.
- The reason is that the variance of both estimators depends on the autocovariances of $\theta_1^{(i)}$ and of $\mathsf{E}(\theta_1|\theta_2^{(i)})$ and not only of the variances like in an IID sample.

Numerical standard error

- If the generated draws are stationary,

$$\text{Var}(g_{mcmc}) = \frac{1}{n}\left(\gamma_0 + 2\sum_{j=1}^{n-1}\gamma_j \frac{n-j}{n}\right) = \frac{S(0)}{n}$$

  where γ_j is the jth-order autocovariance of $\{g(\theta^{(i)}\}_{i=1}^{n}$ and $S(0)$ is the spectral density at 0. This can be estimated consistently.

- The numerical standard error is the square root of this estimated variance divided by n. A probabilistic error bound (or a confidence interval) can be evaluated relying on an asymptotic normality theorem, as for direct sampling and IS.

Subsampling the chain

- Instead of basing the estimate of μ_g on the mean of a single long dependent sequence, with the difficulty of estimating reliably the numerical variance, one can sub-sample the chain, i.e. retain draws that are distant enough from each other such that they can be considered to be independent. The mean of the retained points is an estimator whose numerical standard error is computed like in the case of direct sampling. However, this wastes a lot of points.

Parallel chains

- Another approach is to run the simulation many times, in each case from a different starting value, and to keep one final draw from each simulation (after the warm-up phase).
- These final draws are independent if convergence of the chain has been achieved and the different starting values are well dispersed and drawn independently (e.g. from the prior or an approximation to the posterior).
- An average of the final draws is an estimator of μ_g, whose numerical standard error is computed as in IID sampling.

Convergence diagnostics for MCMC

- Apply Geweke's test to $g(\theta) = \theta$ (element by element).
- Plot and analyze the autocorrelations of the draws.
- Do the same with CUMSUM statistics.
- Useful reference with example of use of the diagnostics: Bauwens, L. and Giot, P. (1998), A Gibbs sampling approach to cointegration, *Computational Statistics* 13, 339-368.
- For a diagnostic based on parallel chains, read Koop's book, pages 67-68.

Geweke's test

- Geweke's test statistic compares the estimate $\bar{g}_A$ of a posterior mean from the first n_A draws with the estimate $\bar{g}_B$ from the last n_B draws. If the two subsamples (of size n_A and n_B) are well separated (i.e. there are many observations between them), they should be independent. The statistic, normally distributed if n is large and the chain has converged, is

  $$Z = \frac{\bar{g}_A - \bar{g}_B}{(nse_A^2 + nse_B^2)^{1/2}}$$

  where nse_A and nse_B are the numerical standard errors of each subsample.

CUMSUM statistics

- The standardized CUMSUM statistic for (scalar) θ is:

  $$CS_t = \left(\frac{1}{t} \sum_{i=1}^{t} \theta^{(i)} - m_\theta \right) / s_\theta,$$

  where m_θ and s_θ are the MC sample mean and standard deviation of the n draws.

- If the MCMC sampler converges, the graph of CS_t against t should converge smoothly to zero. On the contrary, long and regular excursions away from zero are an indication of the absence of convergence.

CUMSUM statistics

- A value of 0.05 for a CUMSUM after t draws means that the estimate of the posterior expectation diverges from the final estimate (after n draws) by 5 per cent in units of the final estimate of the posterior standard deviation; so a divergence of even 25 per cent is not a bad result.
- One could declare that the sampler has converged after $N(\epsilon)$ draws *for the estimation of a certain quantity (like a posterior mean)* with a relative error of $100 \times \epsilon$ per cent, if CS_t remains within a band of $\pm\epsilon$ for all t larger than $N(\epsilon)$. The relative error should be fixed at a low value, such as 0.05.

References

- Chapter 3 of BLR reviews the topics covered in this chapter of the course.
- Other references: see list at the end of Chapter 1, and

  -Specialized books, e.g.:

  Chen, Shao, Ibrahim (2000), Monte Carlo Methods in Bayesian Computation, Springer.

  -Chapters 4 and 26 in Handbook of Computational Statistics (2012, second edition in two volumes), Springer.

COURSE STRUCTURE

- Chapter 1: Concepts (p 2)
- Chapter 2: Numerical Methods (p 33)
- Chapter 3: Single Equation Regression Analysis
- Chapter 4: VAR Models (p 149)

CHAPTER 3

- 3.1 Regression with non-informative prior
- 3.2 Regression with conjugate prior
- 3.3 Partially linear model
- 3.4 Regression with non-conjugate prior
- 3.5 Heteroskedastic errors
- 3.6 Autocorrelated errors
- 3.7 IID Student errors

In 3.1-3.4, error terms are assumed IID $N(0, \sigma^2)$.
In 3.5-3.6, they are still assumed normally distributed.

Gaussian linear regression

- $y_t = x_t'\beta + \epsilon_t,\ \epsilon_t \sim I.N(0,\sigma^2),\ t = 1,\ldots,T$;
  x_t of dimension $k \times 1$ and exogenous for β and σ^2;
  $y = X\beta + u$ in matrix format. Denote $(y\ X)$ by d.
  We assume $T > k$.
- Likelihood function:

  $$L(\beta,\sigma^2|d) \propto (\sigma^2)^{-T/2} \exp[-\tfrac{1}{2}\,\sigma^{-2}(y - X\beta)'(y - X\beta)]$$

  $$= (\sigma^2)^{-T/2} \exp\left(-\tfrac{1}{2}\,\sigma^{-2}[s + (\beta - \hat\beta)'X'X(\beta - \hat\beta)]\right)$$

  where

  $\hat\beta = (X'X)^{-1}X'y$ (the OLS estimator),

  $s = y'M_Xy$ (with $M_X = I_T - X(X'X)^{-1}X'$) is the sum of squared OLS residuals (SSR).

Non-informative prior (NIP)

- It is usually defined as $\varphi(\beta, \sigma^2) \propto 1/\sigma^2$.
- This can be interpreted as uniform on β on $\mathbb{R}^k$, and uniform on $\ln \sigma^2$ on $\mathbb{R}$, since $\varphi(\ln \sigma^2) \propto 1 \Rightarrow \varphi(\sigma^2) \propto 1/\sigma^2$ by the change of variable rule.
- The prior is 'improper': it is not a density, since it does not integrate to 1. However this does not prevent the posterior to be integrable (or proper).
- Instead of saying that a parameter is uniformly distributed on $(-\infty, +\infty)$, we could decide that it should be uniformly distributed on the bounded interval $(-B, +B)$. By choosing B to be very large but finite, the prior is proper and the posterior will be the same as if we use the NIP.

Posterior under NIP

- Multiplying prior and likelihood:

$$\varphi(\beta,\sigma^2|d) \propto (\sigma^2)^{-(T+2)/2} \exp\left[-\frac{(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})+s}{2\sigma^2}\right]$$

- The posterior density is a Normal-inverted-gamma density: NIG$(\hat{\beta}, X'X, s, \nu)$, meaning

$$\begin{aligned}\varphi(\beta,\sigma^2|d) &= \varphi(\beta|\sigma^2,d)\varphi(\sigma^2|d) \quad \text{where}\\ \beta|\sigma^2,d &\sim N_k(\hat{\beta},\sigma^2(X'X)^{-1}) \quad \text{and}\\ \sigma^2|d &\sim IG_2(\nu,s),\end{aligned}$$

where IG_2 means Inverted-Gamma-2, and $\nu = T - k > 0$.

Inverted-Gamma-2 density

- A random variable $X > 0$ has an $IG_2(\nu, s)$ density, where $\nu > 0$, $s > 0$ are the degrees of freedom and scale parameters, if its density is given by

  $$\left(\frac{s}{2}\right)^{\nu/2} \frac{1}{\Gamma(\nu/2)} x^{-\frac{1}{2}(\nu+2)} \exp\left(-\frac{s}{2x}\right).$$

  where $\Gamma(z) = \int_0^z u^{z-1} \exp(-z) dz$.

- $\text{E}(X) = \dfrac{s}{\nu - 2}$ if $\nu > 2$, $\text{Var}(X) = \dfrac{2}{\nu - 4}[\text{E}(X)]^2$ if $\nu > 4$.

- If $X \sim IG_2(\nu, s)$, $Y = 1/X \sim G_2(\nu, s)$. Note that $W \sim G_2(\nu, 1)$ is the same as $W \sim \chi^2(\nu)$. Equivalently, $Y = W/s$ and $X = s/W$.

Posterior under NIP

- The joint posterior density $\varphi(\beta, \sigma^2|d)$ can also be factorized as

$$\begin{aligned} \varphi(\beta, \sigma^2|d) &= \varphi(\beta|d)\varphi(\sigma^2|\beta, d) \quad \text{where} \\ \sigma^2|\beta, d &\sim IG_2(T, (y - X\beta)'(y - X\beta)) \\ \beta|d &\sim t_k(\hat{\beta}, s, X'X, \nu). \end{aligned}$$

- This shows that the marginal posterior density of β is a multivariate t (Student) density, with parameters $\hat{\beta}$, $X'X$, s and ν.

Multivariate t-density

- A random vector $X \in \mathbb{R}^k$ has a Student (or t) distribution with parameters $\nu > 0$ (degrees of freedom), $\mu \in R^k$, M a positive-definite matrix of order k, and $s > 0$, i.e. $X \sim t_k(\mu, s, M, \nu)$, if its density function is given by

$$f_t(x|\mu, s, m, \nu) = \frac{\Gamma(\frac{\nu+k}{2})}{\Gamma(\frac{\nu}{2})\pi^{\frac{k}{2}}} s^{\frac{1}{2}\nu} |M|^{\frac{1}{2}} [s+m\,(x-\mu)' M (x-\mu)]^{-\frac{1}{2}(\nu+k)}.$$

- Its mean and variance-covariance matrix are

$$\mathsf{E}(X) = \mu \ \text{ if } \nu > 1, \quad \mathsf{Var}(X) = \frac{s}{\nu-2} M^{-1} \ \text{ if } \nu > 2.$$

Posterior moments under NIP

- Analytical results are available, such as

$$\begin{aligned} \mathsf{E}(\beta|d) &= \hat{\beta}, \\ \mathsf{Var}(\beta|d) &= \frac{s}{\nu-2}(X'X)^{-1} = \mathsf{E}(\sigma^2|d)(X'X)^{-1}, \end{aligned}$$

- Direct sampling can be used: e.g. if we are interested by the marginal posterior density of $(\beta_1+\beta_2)/(1-\beta_3)$, we

  1. *Generate* R *draws* $\{\beta^{(r)}\}_{r=1}^{R}$ *of* β *from* $t_k(\hat{\beta}, s, X'X, \nu)$.
  2. *Compute* $(\beta_1^{(r)}+\beta_2^{(r)})/(1-\beta_3^{(r)})$ *for* $r = 1, 2, \ldots, R$.
  3. *Use a kernel method to estimate of the posterior density.*

  NB: moments of such a ratio do not exist!

CHAPTER 3

- 3.1 Regression with non-informative prior
- 3.2 Regression with conjugate prior
- 3.3 Partially linear model
- 3.4 Regression with non-conjugate prior
- 3.5 Heteroskedastic errors
- 3.6 Autocorrelated errors
- 3.7 IID Student errors

In 3.1-3.4, error terms are assumed IID $N(0, \sigma^2)$.
In 3.5-3.6, they are still assumed normally distributed.

Conjugate prior densities

- A (natural) conjugate prior (CP) density has the same functional form (with respect to θ) as the likelihood. The posterior density then retains the same functional form as the prior.
- CP densities provide analytical posterior results: no need for numerical integration!
- Drawback: they restrict the class of prior densities.
- They are useful tools for more complex models.
- They exist when the data density belong to the exponential family, so that sufficient statistics exist. See BLR, sections 2.4 and 2.5 for details.

CP for linear regression

- The likelihood has the functional form of a NIG$(\hat{\beta}, X'X, s, T-k-2)$ density for $(\beta' \; \sigma^2)$.
- Hence, the CP class is NIG$(\beta_0, M_0, s_0, \nu_0)$, where $s_0 > 0$, $\nu_0 > 0$, $\beta_0 \in R^k$ and M_0 is a PDS matrix $(k \times k)$.
- Written explicitly, the prior kernel is

  $$(\sigma^2)^{-(\nu_0+k+2)/2} \exp\left(-\tfrac{1}{2}\,\sigma^{-2}[s_0 + (\beta-\beta_0)' M_0 (\beta-\beta_0)]\right).$$

- Some prior moments:

  E$(\beta) = \beta_0$, Var$(\beta) = M_0^{-1} s_0/(\nu_0 - 2)$ if $\nu_0 > 2$,

  E$(\sigma^2) = s_0/(\nu_0 - 2)$.

Posterior under CP

- The posterior is NIG$(\beta_*, M_*, s_*, \nu_*)$, where

  $$M_* = M_0 + X'X$$

  $$\beta_* = M_*^{-1}(M_0\beta_0 + X'X\hat{\beta})$$

  $$s_* = s_0 + s + (\beta_0 - \hat{\beta})'[M_0^{-1} + (X'X)^{-1}]^{-1}(\beta_0 - \hat{\beta})$$

  $$\nu_* = \nu_0 + T$$

  $$\Rightarrow$$

- $\beta|d \sim t_k(\beta_*, s_*, M_*, \nu_*)$ and

  E$(\beta|d) = \beta_*$, Var$(\beta|d) = s_*/(\nu_* - 2)M_*^{-1}$.

- $\sigma^2|d \sim IG_2(\nu_*, s_*)$ and E$(\sigma^2) = s_*/(\nu_* - 2)$.

- $\beta|\sigma^2, d \sim N_k(\beta_*, \sigma^2 M_*^{-1})$ and

  $\sigma^2|\beta, d \sim IG_2(\nu_* + k, s_* + (\beta - \beta_*)'M_*(\beta - \beta_*))$.

Gibbs sampling for posterior under CP

For computing special posterior features, one can use direct sampling (as for the NIP). Another option is a Gibbs sampling algorithm to generate R draws from the posterior of β and σ^2 (after R_0 warming-up draws):

1. *Choose an initial value* $(\sigma^2)^{(0)}$ *(e.g.* $SSR/(T-k)$*).*
2. *Set* $r = 1$.
3. *Draw successively* $\beta^{(r)}$ *from* $N_k\left(\beta_*, (\sigma^2)^{(r-1)} M_*^{-1}\right)$ and $(\sigma^2)^{(r)}$ *from* $IG_2(\nu_* + k, s_* + (\beta^{(r)} - \beta_*)' M_* (\beta^{(r)} - \beta_*))$.
4. *Set* $r = r + 1$ *and go to step 3 unless* $r > R_0 + R$.
5. *Discard the first* R_0 *values of* $\beta^{(r)}$ *and* $(\sigma^2)^{(r)}$*. Compute what you are interested in from the last* R *draws.*

Proof

- It can be checked directly that

  $s_0 + (\beta - \beta_0)' M_0 (\beta - \beta_0) + s + (\beta - \hat\beta)' X' X (\beta - \hat\beta)$

  $= (\beta - \beta_*)' M_* (\beta - \beta_*) + s_0 + s + \beta_0' M_0 \beta_0 + \hat\beta' X' X \hat\beta - \beta_*' M_* \beta_*$,

  so that

  $s_* = s_0 + s + \beta_0' M_0 \beta_0 + \hat\beta' X' X \hat\beta - \beta_*' M_* \beta_*$.

  More algebra allows to express s_* as on page 107.
- Then the posterior kernel is

  $(\sigma^2)^{-(T+\nu_0+k+2)/2} \exp\left(-\frac{1}{2}\,\sigma^{-2}[s_* + (\beta - \beta_*)' M_* (\beta - \beta_*)]\right)$,

  which has the form of a NIG density kernel.

NIP and CP

- The NIP prior $\propto 1/\sigma^2$ is the limit of the NIG conjugate prior kernel obtained when$M_0 \to 0$, $s_0 \to 0$, $\nu_0 \to -k$.
- These values are at the boundary of the admissible values of $M_0 > 0$, $s_0 > 0$.
  For ν_0, the limit is $-k$ rather than 0 (sometimes used) because then $\nu_* = T - k$ (instead of T): it is sensible to have a degree of freedom correction.
- The value of β_0 is irrelevant when $M_0 = 0$ and is fixed at $\beta_0 = 0$ for simplicity.
- With $M_0 = 0$, $s_0 = 0$, $\nu_0 = -k$, and $\beta_0 = 0$, the posterior under the CP is the same as the posterior under the NIP.

Partially non-informative conjugate prior

- Let $\beta' = (\beta_1' \; \beta_2')$: one can be non-informative on e.g. β_2 and use a conjugate prior on β_1, by setting

$$\beta_0 = \begin{pmatrix} \beta_{0,1} \\ 0 \end{pmatrix}, \qquad M_0 = \begin{pmatrix} M_{0,11} & 0 \\ 0 & 0 \end{pmatrix}.$$

  NB: M_0^{-1} is like M_0 but with $M_{0,11}^{-1}$.

- The posterior parameters given three pages above are still well defined.

Prior elicitation of β/1

- For each element of β on which one has prior information, assign directly the prior mean and standard deviation of the Student distribution.

  Given the values of s_0 and $\nu_0 > 2$ chosen for the IG_2 prior of σ^2, deduce the values of β_0 and M_0:

$$\beta_0 = \mathsf{E}(\beta), \qquad M_0 = [\mathsf{Var}(\beta)]^{-1} s_0/(\nu_0 - 2).$$

- Example: given $\mathsf{E}(\beta_j)$ and $\mathsf{Var}(\beta_j) \Rightarrow \beta_{0,j} = \mathsf{E}(\beta_j)$ and $m_{0,jj} = [1/\mathsf{Var}(\beta_j)] s_0/(\nu_0 - 2)$.
- The latter is correct if $\mathsf{Var}(\beta)$ is DIAGONAL, as is often assumed for simplicity.

Prior elicitation of β/2

- Prior beliefs may come from theoretical restrictions or previous empirical results on similar (but different) data.
- Example 1: theory constrains that $\beta_L \leq \beta \leq \beta_U$ (β being scalar). The prior mean can be $0.5(\beta_L + \beta_U)$. If the marginal prior is close to a Normal, fixing the prior standard deviation to $(\beta_U - \beta_S)/6$, implies that Pr$(\beta_L \leq \beta \leq \beta_U) \approx 1$.
- Example 2: let β be the AR(1) coefficient of a dynamic regression for the inflation rate. It should be less than 1 (inflation is not explosive), and it is likely to be positive (there is some persistence in inflation). One could assume that $\beta_L = 0$ and $\beta_U = 0.8$ and proceed as above. A less informative prior is based on $\beta_L = -1$ and $\beta_U = 1$.

Prior elicitation of β/3

- An annoying aspect of the CP is that $\text{Var}(\beta|\sigma^2) = \sigma^2 M_0^{-1}$. Hence, we cannot just fix $\text{Var}(\beta)$ to find M_0 since $M_0 = [\text{Var}(\beta)]^{-1} s_0/(\nu_0 - 2)$.
- We must choose also s_0 and ν_0, i.e. we must elicit a prior for σ^2.
- It is not easy to have prior beliefs about σ^2: we would prefer to be non-informative on it. But setting $s_0 = 0$ implies that $M_0 = [\text{Var}(\beta)]^{-1} s_0/(\nu_0 - 2)$ is then equal to 0: we would be also non-informative on β_0.
- To be practical, we can fix $s_0 = s$ (the SSR), choose ν_0 (e.g. $= 3$) and deduce M_0.
- To avoid such arbitrary tricks, we should use a non-conjugate prior.

Zellner's g-prior

- Zellner's proposed the following CP:

$$\beta|\sigma^2 \sim N_k(0, g\sigma^2(X'X)^{-1})$$

  where $g > 0$ is a scalar value to be chosen.

- The g-prior mean is $\beta_0 = 0$. The posterior mean is

$$\bar{\beta} = \frac{g}{1+g}\hat{\beta}.$$

  If $g \to \infty$, $\bar{\beta} \to \hat{\beta}$, and if $g \to 0$, $\bar{\beta} \to 0$.

- The posterior variance is

$$\text{Var}(\beta|d) = \text{E}(\sigma^2|d)\frac{g}{1+g}(X'X)^{-1}.$$

CHAPTER 3

- 3.1 Regression with non-informative prior
- 3.2 Regression with conjugate prior
- 3.3 Partially linear model
- 3.4 Regression with non-conjugate prior
- 3.5 Heteroskedastic errors
- 3.6 Autocorrelated errors
- 3.7 IID Student errors

In 3.1-3.4, error terms are assumed IID $N(0, \sigma^2)$.
In 3.5-3.6, they are still assumed normally distributed.

Posterior under CP: Another presentation

- Model: $y = X\beta + \epsilon$, $\epsilon \sim N_T(0, \sigma^2 I_T)$.
- Prior: $\beta|\sigma^2 \sim N_k(\beta_0, \sigma^2 M_0^{-1})$. This can be expressed as: $\beta = \beta_0 + u$ with $u \sim N_k(0, \sigma^2 M_0^{-1})$ or as $\beta_0 = \beta - u$, so that

$$(M_0^{1/2})'\beta_0 = (M_0^{1/2})'\beta + \epsilon_0,\ \epsilon_0 = -(M_0^{1/2})'u \sim N_k(0, \sigma^2 I_k).$$

- Stacking, we get an extended regression:

$$\begin{pmatrix} y \\ (M_0^{1/2})'\beta_0 \end{pmatrix} = \begin{pmatrix} X \\ (M_0^{1/2})' \end{pmatrix} \beta + \begin{pmatrix} \epsilon \\ \epsilon_0 \end{pmatrix} \text{ with } \begin{pmatrix} \epsilon \\ \epsilon_0 \end{pmatrix} \sim N(0, \sigma^2 I_{T+k}).$$

- Combining the prior $\sigma^2 \sim IG_2(\nu_0, s_0)$ and $\varphi(\beta) \propto 1$ with the likelihood function of the extended regression gives the posterior defined on page 107: in particular, β_* is the OLS formula and s_* the OLS SSR applied to this extended regression.

Partially linear model

- Consider $y_t = x_t'\beta + f(z_t) + \epsilon_t$, where z_t *is a scalar* and $f(.)$ an unknown function. Assume $\epsilon_t \sim I.N(0, \sigma^2)$. The vector x_t should not contain a 1 since $f(z_1)$ plays the role of the intercept.
- Observations must be ordered like $z_1 \leq z_2 \leq \ldots < \ldots \leq z_T$, which may not be sensible, e.g. in time series data (except if $z_t = t$).
- Notations: $\gamma = (f(z_1)\ f(z_2) \ldots f(z_T))'$, $W = (X\ I_T)$, $\delta = (\beta'\ \gamma')'$. Then $y = W\delta + \epsilon$.
- There are more coefficients than observations ($k + T > T$). We need more information to overcome the fact that $W'W$ is singular.
  A non-informative prior is excluded!

What prior information?

- A smoothing prior: the function $f(z_t)$ is likely to be smooth, e.g. such that $\gamma_t - \gamma_{t-1}$ is small for all t. This fixes the prior expectations of the differences at 0. Then one can fix the prior variances at the same small value η, and the prior covariances at 0.
- Defining $P_0^{-1} = \mathrm{diag}(\eta\,\eta \ldots \eta)$, the prior on γ can be formalized as follows: $D\gamma \sim N_{T-1}(0, \sigma^2 P_0^{-1})$, where D is the first-differencing matrix, of dimension $T-1 \times T$, such that $D\gamma = (\gamma_2 - \gamma_1,\ \gamma_3 - \gamma_2 \ldots \gamma_T - \gamma_{T-1})'$.
- Combined with an IG_2 prior on σ^2, the prior on $D\gamma$ is conjugate. It can be extended to include a conjugate prior on β.

Posterior

- Even if the prior is non-informative on β, the posterior is integrable, i.e. $W'W + M_0$ is of full rank, where $M_0 = \begin{pmatrix} 0 & 0 \\ 0 & P_0 \end{pmatrix}$.
- The extended regression is: $\begin{pmatrix} y \\ 0_{T-1} \end{pmatrix} = \begin{pmatrix} X & I_T \\ 0 & (P_0^{1/2})'D \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} \epsilon \\ \epsilon_0 \end{pmatrix}$, and the parameters of the NIG posterior can be computed as explained for the conjugate prior.
- For more details, see Koop and Poirier (2004), Bayesian variants of classical semiparametric regression techniques, Journal of Econometrics.

CHAPTER 3

- 3.1 Regression with non-informative prior
- 3.2 Regression with conjugate prior
- 3.3 Partially linear model
- 3.4 Regression with non-conjugate prior
- 3.5 Heteroskedastic errors
- 3.6 Autocorrelated errors
- 3.7 IID Student errors

In 3.1-3.4, error terms are assumed IID $N(0, \sigma^2)$.
In 3.5-3.6, they are still assumed normally distributed.

Normal-diffuse prior

- A drawback of the conjugate prior is that one has to be informative on σ^2 to be informative on β, whereas one may prefer, for simplicity, to use $\varphi(\sigma^2) \propto 1/\sigma^2$.
- To avoid this, one can define the prior as independent between β and σ^2, with $\varphi(\sigma^2)$ non-informative. Then a convenient prior for β is a Normal prior: $\beta \sim N_k(\beta_0, M_0^{-1})$.
- Notice that $\mathsf{Var}(\beta) = M_0^{-1}$ does not depend on σ^2 as in the CP case.
- Elicitation of M_0 can be done directly without having to bother about s_0 and ν_0 as in the CP case: now

  $M_0 = [\mathsf{Var}(\beta)]^{-1}$.

Posterior under Normal-diffuse prior

- The posterior $\varphi(\beta, \sigma^2|d)$ is proportional to $(\sigma^2)^{-(T+2)/2} \exp\left(-\frac{1}{2}\,\sigma^{-2}[s + (\beta - \hat{\beta})' X' X (\beta - \hat{\beta})]\right)$ $\times \exp\left[-0.5(\beta - \beta_0)' M_0 (\beta - \beta_0)\right]$.
- Simple computations show that

  1) $\beta|\sigma^2, d \sim N_k(\overline{\beta}^*, \overline{V}^*)$, where
  $\overline{V}^* = \left(M_0 + \sigma^{-2} X' X\right)^{-1}$,
  $\overline{\beta}^* = \overline{V}^* \left(M_0 \beta_0 + \sigma^{-2} X' X \hat{\beta}\right)$,
  2) $\sigma^2|\beta, d \sim IG_2(T, (y - X\beta)'(y - X\beta))$.
- Computations of posterior features can easily be done using a Gibbs sampler similar to the one for the CP case, but using the conditional densities defined above.

Special cases of Normal prior/1

- Ridge regression prior: $\beta \sim N_k(0, \tau I_k)$. It leads to a posterior mean similar to the estimate obtained from classical ridge regression: $\overline{\beta}^* = \left(\sigma^{-2}X'X + \tau^{-1}I_k\right)^{-1}\sigma^{-2}X'y$.
- If $\tau \to \infty$, $\overline{\beta}^* \to \hat{\beta}$, and if $\tau \to 0$, $\overline{\beta}^* \to 0$, like with the g-prior. However, for $0 < \tau < \infty$, the Gibbs sampler must be used.
- How to choose τ is a sensitive issue.
- A drawback is that the same amount of shrinkage is imposed on all coefficients.

Special cases of Normal prior/2

- Hierarchical shrinkage prior: use independent Normal priors on each element of β: $\beta_i \sim N(0, \tau_i)$ for $i = 1, 2, \ldots, k$.
- To avoid choosing τ_i for each coefficient, we use a common prior for them : $\tau_i \sim IG_2(q_1, q_2)\, \forall i$.
- The posterior can be simulated by a Gibbs sampler:

1. Draw $\tau_i|\beta_i$ from $IG_2\left(q_1 + 1, q_2 + \beta_i^2\right)\ \forall i$.
2. Draw $\sigma^2|\beta, d$ from $IG_2\left(T, (y - X\beta)'(y - X\beta)\right)$.
3. Draw $\beta|\tau_1, \ldots, \tau_k, \sigma^2, d$ from

$$N_k\left(\left(\sigma^{-2}X'X + \underline{V}^{-1}\right)^{-1} X'y, \left(\sigma^{-2}X'X + \underline{V}^{-1}\right)^{-1}\right)$$

where $\underline{V} = \mathrm{diag}\left(\tau_1, ..., \tau_k\right)$ is the prior covariance matrix of β given $\tau_1, \tau_2, \ldots, \tau_k$.

Non-normal-diffuse prior

- Instead of the $N_k(\beta_0, M_0^{-1})$ prior, one can use any other prior $\varphi(\beta)$. This is necessary to avoid using a symmetric prior for a parameter.
- Example: the long-run marginal propensity to consume in a macro equation should be smaller than but close to 1. A symmetric prior should have a very small variance. An asymmetric prior can avoid this.
- It could be a Beta prior or a truncated Normal prior. The latter is convenient since the previous results for the Normal-diffuse prior can be used, adding the indicator function for the prior on β to the prior and posterior. Then a rejection step must be added to the Gibbs sampler: reject any draw of β that does not lie in the required region.

MH step in Gibbs sampler

- In general $\varphi(\beta|\sigma^2, d) \propto \varphi(\beta) \exp\left(-\frac{1}{2}(\beta - \hat{\beta})' \frac{X'X}{\sigma^2}(\beta - \hat{\beta})\right)$.
  Hence $\beta|\sigma^2, d$ is not Normal and not in a known class.
- Then drawing directly $\beta|\sigma^2, d$ in a Gibbs sampler is not feasible: one can draw using a MH step.
- The proposal density can be designed as follows:
  -approximate the prior $\varphi(\beta)$ by a Normal density;
  -compute the posterior under the approximating Normal prior, say $N_k(\overline{\beta}_*, \overline{V}^*)$ (4 pages above).
  -use as proposal for the MH step this Normal density.
- If the Normal approximation is not good, the MH step will be inefficient.

CHAPTER 3

- 3.1 Regression with non-informative prior
- 3.2 Regression with conjugate prior
- 3.3 Partially linear model
- 3.4 Regression with non-conjugate prior
- 3.5 Heteroskedastic errors
- 3.6 Autocorrelated errors
- 3.7 IID Student errors

In 3.1-3.4, error terms are assumed IID $N(0, \sigma^2)$.
In 3.5-3.6, they are still assumed normally distributed.

Heteroskedasticity

- $y_t = x_t'\beta + \epsilon_t,\ \epsilon_t \sim I.N(0, \sigma_t^2),\ t = 1, \ldots, T$ where

  $\sigma_t^2 = \sigma^2\, h(z_t, \alpha) > 0$

  z_t: vector of ℓ variables, may be functions of x_t,
  but NOT of $\beta \Rightarrow$ GARCH models excluded!

  α: vector of ℓ parameters

  no constant term in z_t (its role is taken by σ^2)

  $h(.)$ defined so that $h(z_t, 0) = 1$

  x_t and z_t weakly exogenous for β, σ^2 and α.

- Examples:

  $h(z_t'\alpha) = \exp(z_t'\alpha), \alpha \in \mathbb{R}^\ell$.

  $h(z_t'\alpha) = 1 + z_t'\alpha$, $\alpha \in A \subset \mathbb{R}^\ell$ such that
  $1 + z_t'\alpha > 0,\ \forall z_t$.

Likelihood function

- Let $H(\alpha) = \text{diag}(1/h(z_1'\alpha),\, 1/h(z_2'\alpha),\, \ldots,\, 1/h(z_T'\alpha))$.
- Likelihood function:

$$
\begin{aligned}
& L(\beta, \sigma^2, \alpha; d) \\
& \propto \sigma^{-T} \sqrt{|H(\alpha)|} \exp\left[-\frac{1}{2\,\sigma^2}(y - X\beta)' H(\alpha)(y - X\beta)\right]
\end{aligned}
$$

$$
\propto \sigma^{-T} \sqrt{|H(\alpha)|} \exp\left(-\frac{1}{2\,\sigma^2}\{[\beta - b(\alpha)]' X' H(\alpha) X [\beta - b(\alpha)] + s(\alpha)\}\right),
$$

where

$$
\begin{aligned}
b(\alpha) &= [X' H(\alpha) X]^{-1} X' H(\alpha) y, \\
s(\alpha) &= y' [H(\alpha) - H(\alpha) X (X' H(\alpha) X)^{-1} X' H(\alpha)] y.
\end{aligned}
$$

Prior density

- Conditionally on α, the model can be transormed into a homosckedastic one, and results for this case can be applied.
- For this factorize the prior as

  $$\varphi(\beta, \sigma^2, \alpha) = \varphi(\beta|\sigma^2)\,\varphi(\sigma^2)\,\varphi(\alpha),$$

  and choose a NIG prior density for σ^2 and β:
  $\sigma^2 \sim IG_2(\nu_0, s_0), \quad \beta|\sigma^2 \sim N(\beta_0, \sigma^2 M_0^{-1})$,
  or a Normal for β times a non-informative prior for σ^2.
- The prior on α can be chosen as one wishes since numerical integration wrt this parameter must be used. An easy to use flat prior is $\varphi(\alpha) \propto 1$ if $\alpha \in A$.

Posterior density of β and $\sigma^2|\alpha$

- If the prior is NIG:

  $\beta|\alpha, d \sim t_k(\beta_*(\alpha), s_*(\alpha), M_*(\alpha), \nu_*)$

  $\sigma^2|\alpha, d \sim IG_2(\nu_*, s_*(\alpha)),$ where

$$\begin{aligned} M_*(\alpha) &= M_0 + X'H(\alpha)X \\ \beta_*(\alpha) &= M_*^{-1}(\alpha)[M_0\beta_0 + X'H(\alpha)y] \\ s_*(\alpha) &= s_0 + s(\alpha) + b'(\alpha)X'H(\alpha)Xb(\alpha) - \beta_*(\alpha)'M_*(\alpha)\beta_*(\alpha) \\ \nu_* &= \nu_0 + T. \end{aligned}$$

- Proof: multiply the likelihood and the prior of β and σ^2, express this product as the kernel of a NIG density and apply the properties of the NIG.

Marginal posterior densities

- Method: integrate likelihood times prior wrt to β and σ^2. Since this product depends on β and σ^2 through the NIG kernel, the result is:
  the integral of the NIG kernel × $|H(\alpha)|^{1/2}$ (from the likelihood function) × the prior density of α:

$$\varphi(\alpha|d) \propto |H(\alpha)|^{1/2}\,|M_*(\alpha)|^{-1/2}\,s_*(\alpha)^{-(\nu_*-k)/2}\,\varphi(\alpha).$$

- Then one can marginalize $\beta|\alpha, d$ and its moments wrt to α:
  $\varphi(\beta|d) \propto \int \varphi(\beta|\alpha, d)\kappa(\alpha|d)d\alpha$, and
  $\mathsf{E}(\beta|d) = \int \beta_*(\alpha)\kappa(\alpha|d)d\alpha / \int \kappa(\alpha|d)d\alpha$ since
  $$\beta_*(\alpha) = \mathsf{E}(\beta|\alpha, d).$$
  Same method for getting $\sigma^2|d$ and its moments.

How to integrate?

- If ℓ (dimension of α) ≤ 2: deterministic integration.
  If $1 < \ell < 10$, griddy-Gibbs on $\alpha|d$. Very convenient if constraints on α.
  If $\ell \geq 10$, importance sampling or MH sampling. Difficult if constraints on α.
- NB: since $\varphi(\alpha|\beta, \sigma^2, d)$ is not better known than $\varphi(\alpha|d)$, a Gibbs sampler cycling between $\alpha|\beta, \sigma^2$, $\beta|\alpha, \sigma^2$, $\sigma^2|\beta, \alpha$ is not particularly interesting!
- For an application: see BLR, p 202-203 (see next page).
- For GARCH models: see sections 7.3 and 7.4, and paper of Bauwens and Lubrano (1998) in the *Econometrics Journal*.

Application

- y_t = household t electricity consumption,
x_t = constant, 10 dummies for ownership of specific electric appliances, socio-economic variables (income, household size, house size...), and interaction variables.
- Data for 174 households.
- Coefficient of a dummy measures electricity consumption due to corresponding appliance. Should be positive: positivity constraints on dummy coefficients.
- Heteroskedasticity due to variability of number of appliances z_t owned by each household: $\sigma_t^2 = \sigma^2 z_t^\alpha$.

CHAPTER 3

- 3.1 Regression with non-informative prior
- 3.2 Regression with conjugate prior
- 3.3 Partially linear model
- 3.4 Regression with non-conjugate prior
- 3.5 Heteroskedastic errors
- 3.6 Autocorrelated errors
- 3.7 IID Student errors

In 3.1-3.4, error terms are assumed IID $N(0, \sigma^2)$.
In 3.5-3.6, they are still assumed normally distributed.

Autocorrelation

- $y_t = x_t'\beta + u_t,\ \rho(L)u_t = \epsilon_t \sim I.N(0,\sigma^2),\ t = 1,\ldots,T+p$ where $\rho(L) = 1 - \rho_1 L - \rho_2 L^2 - \ldots - \rho_p L^p$.
- We assume that p initial observations ($y_{1-p} \ldots y_{-1}\ y_0$) are used as initial conditions.
- Equivalently: $\rho(L)y_t = \rho(L)x_t'\beta + \epsilon_t$.

  The model is non-linear in the parameters.

  For example, with x_t scalar and $\rho(L) = 1 - \rho L$, this is

  $y_t = \rho y_{t-1} + x_t\beta + x_{t-1}\rho\beta + \epsilon_t$.
- Let $\rho = (\rho_1\ \rho_2 \ldots \rho_p)$. Given ρ, the model is linear in β, and given β it is linear in ρ. In each case, one can apply the results for linear regression with a conjugate prior.

Prior and posterior

- Hence a convenient way to factorize the prior is

$$\varphi(\beta, \rho, \sigma^2) = \varphi(\beta|\sigma^2)\ \varphi(\rho|\sigma^2)\ \varphi(\sigma^2),$$

  where $\beta|\sigma^2 \sim N_k(\beta_0, \sigma^2 M_0^{-1})$, $\rho|\sigma^2 \sim N_p(\rho_0, \sigma^2 P_0^{-1})$, and $\sigma^2 \sim IG_2(\nu_0, s_0)$. This implies that $(\beta,\ \rho,\ \sigma^2) \sim NIG$.

- In this setup, we can show that:

$$\begin{aligned} \beta|\rho, \sigma^2, d &\sim N_k(\beta_*(\rho), \sigma^2 M_*(\rho)^{-1}) \\ \rho|\beta, \sigma^2, d &\sim N_p(\rho_*(\beta), \sigma^2 P_*(\beta)^{-1}), \\ \sigma^2|\beta, \rho, d &\sim IG_2(\nu_*, s_*(\beta, \rho)). \end{aligned}$$

  so that a Gibbs sampler can be applied.

Details

- For example, to obtain $\beta|\rho, \sigma^2, d$, write the model as $y_t(\rho) = x_t(\rho)'\beta + \epsilon_t$ where $y_t(\rho) = \rho(L)y_t$ and $x_t(\rho) = \rho(L)x_t$, and apply the formulas for the homoskedastic linear regression model. Hence, $M_*(\rho) = M_0 + X(\rho)'X(\rho)$ $\beta_*(\rho) = M_*(\rho)^{-1}[M_0\beta_0 + X(\rho)'y(\rho)]$,
  where $X(\rho)$ is the matrix with $x_t(\rho)'$ as t-th row, and $y(\rho)$ is the vector with $y_t(\rho)$ as t-th element.
- Likewise, to obtain $\rho|\beta, \sigma^2, d$, write the model as $y_t(\beta) = x_t(\beta)'\rho + \epsilon_t$ where $y_t(\beta) = y_t - x_t'\beta$ and $x_t(\beta) = [y_{t-1}(\beta)\ y_{t-2}(\beta) \ldots y_{t-p}(\beta)]$.
  Stack $y_t(\beta)$ in the vector $y(\beta)$ and $x_t(\beta)'$ in $X(\beta)$, and get $P_*(\beta)$ and $\rho_*(\beta)$.

Remarks

- If $\rho(L)$ has a unit root, i.e. $\rho(1) = 0$, the matrix $X(\rho)'X(\rho)$ is singular. This creates a problem in the Gibbs sampler when ρ is close to the value $\rho(1) = 0$. Solutions:
  -be informative on the constant term;
  -work with data in deviation from means;
  -exclude a priori $\rho(1)$ smaller than a threshold.
- The prior may incorporate an indicator function that ρ is in the region of stationarity. Then the posterior is truncated to that region. In the Gibbs algorithm, draws that are not in the region of stationary values should be rejected.

CHAPTER 3

- 3.1 Regression with non-informative prior
- 3.2 Regression with conjugate prior
- 3.3 Partially linear model
- 3.4 Regression with non-conjugate prior
- 3.5 Heteroskedastic errors
- 3.6 Autocorrelated errors
- 3.7 IID Student errors

In 3.1-3.4, error terms are assumed IID $N(0, \sigma^2)$.
In 3.5-3.6, they are still assumed normally distributed.

IID Student errors

- $y_t = x_t'\beta + \epsilon_t,\ \epsilon_t \sim I.t(0, 1, \nu^{-1}\sigma^{-2}, \nu),\ t = 1, \ldots, T$;
  E$(\epsilon_t) = 0$ if $\nu > 1$, Var$(\epsilon_t) = \nu\sigma^2/(\nu - 2)$ if $\nu > 2$.
  If $\nu \to \infty$,we are back to the Gaussian case.
  If ν is small, we have thick tails.
- Likelihood function:
  $L(\beta, \sigma^2, \nu|d) \propto \prod_{t=1}^{T} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \left(\nu\sigma^2\right)^{-1/2} \left[1 + \frac{(y_t - x_t'\beta)^2}{\nu\sigma^2}\right]^{-(\nu+1)/2}$.
- A simple prior: $\varphi(\beta, \sigma^2, \nu) = \varphi(\beta|\sigma^2)\varphi(\sigma^2)\varphi(\nu)$, with $\varphi(\beta) \propto 1$ or $N(\beta_0, \sigma^2 M_0^{-1})$, $\varphi(\sigma^2)$ an $IG_2(\nu_0, s_0)$ and $\varphi(\nu)$ to be specified.
- Posterior not so simple! Full conditional densitites $\varphi(\beta|\sigma^2, \nu, d)$, $\varphi(\sigma^2|\beta, \nu, d)$, $\varphi(\nu|\beta, \sigma^2, d)$ are not known. MH steps within Gibbs do not seem easy.

A solution: data augmentation

- Write $y_t = x_t'\beta + \lambda_t u_t,\ u_t \sim I.N(0, \sigma^2)$, and assume $\lambda_t^2 \sim I.IG_2(\nu, \nu)$ and independent of u_t. Then: $\lambda_t u_t | \lambda_t \sim N(0, \sigma^2 \lambda_t^2)$ and $\lambda_t^2 \sim IG_2(\nu, \nu)$ imply that $\lambda_t u_t \sim t(0, \nu, \sigma^{-2}, \nu) \equiv t(0, 1, \nu^{-1}\sigma^{-2}, \nu)$. Furthermore, $\{\epsilon_t = \lambda_t u_t\}$ is an independent sequence.
- Let $\lambda = (\lambda_1\ \lambda_2 \ldots \lambda_T)$. We consider that β, σ^2, ν and λ are the parameters of the model. Then a Gibbs sampler is feasible, cycling between
  $\beta | \sigma^2, \nu, \lambda, d \sim N_k(\beta_*(\lambda), \sigma^2 M_*(\lambda)^{-1})$,
  $\sigma^2 | \beta, \nu, \lambda, d \sim IG_2(\nu_*, s_*(\beta, \lambda))$,
  $\nu | \beta, \sigma^2, \lambda, d$ a density that can be simulated,
  $\lambda | \beta, \sigma^2, \nu, d \sim \prod_{t=1}^{T} IG_2()$ (hence each λ_t can be simulated independently of the other).

λ known

- Write $y_t/\lambda_t = (x_t'/\lambda_t)\beta + u_t$. This is a homoskedastic Gaussian regression model. Stacking the y_t/λ_t in the vector y_λ and the x_t'/λ_t in the matrix X_λ and applying the results for the Gaussian regression under a conjugate prior gives the conditional densities defined on the previous slide, with:

  $M_*(\lambda) = M_0 + X_\lambda' X_\lambda$
  $\beta_*(\lambda) = M_*(\lambda)^{-1}(M_0\beta_0 + X_\lambda' X_\lambda \hat{\beta}_\lambda)$
  $\nu_* = \nu_0 + T \; s_*(\beta, \lambda) = s_*(\lambda) + [\beta - \beta_*(\lambda)]' M_*(\lambda)[\beta - \beta_*(\lambda)]$
  where $\hat{\beta}_\lambda = (X_\lambda' X_\lambda)^{-1} X_\lambda' y_\lambda$, and
  $s_*(\lambda) = s_0 + y_\lambda' M_{X_\lambda} y_\lambda + \beta_0' M_0 \beta_0 + \hat{\beta}_\lambda' X_\lambda' X_\lambda \hat{\beta}_\lambda - \beta_*'(\lambda) M_*(\lambda) \beta_*(\lambda)$.
- Note that if λ is known, nothing depends on ν.

Full conditional of λ

- Write the model as $(y_t - x_t'\beta)/\sigma = \tilde{y}_t = \lambda_t \tilde{u}_t$ where $\tilde{u}_t \sim I.N(0,1)$. Stack the $\tilde{y}_t$ in the vector $\tilde{y}$.
  The likelihood function, proportional to the joint density of $\tilde{y}$ and λ, is built as
  $\prod_{t=1}^{T} f(\tilde{y}_t|\lambda_t^2) f(\lambda_t^2) \propto$
  $\prod_{t=1}^{T} (\lambda_t^2)^{-1/2} \exp\left(-\frac{\tilde{y}_t^2}{2\lambda_t^2}\right) (\lambda_t^2)^{-(\nu+2)/2} \exp\left(-\frac{\nu}{2\lambda_t^2}\right) =$
  $\prod_{t=1}^{T} (\lambda_t^2)^{-(\nu+3)/2} \exp\left(-\frac{\nu+\tilde{y}_t^2}{2\lambda_t^2}\right)$
- Viewing this function as a function of λ given the data and the other parameters, it is clear that this is a product of independent $IG_2(\nu+1, \nu+\tilde{y}_t^2)$ densities, one for each λ_t.

Full conditional of ν

- It is equal to the prior $\varphi(\nu)$, times the likelihood of $\tilde{y}$ and λ, where we keep only all that depends on ν: $\varphi(\nu|\beta,\sigma^2,\lambda,d) \propto$ $\varphi(\nu)\prod_{t=1}^{T}\left[\frac{1}{\Gamma(\nu/2)}\nu^{\nu/2}\left(\lambda_t^2\right)^{-(\nu+2)/2}\exp\left(-\frac{\nu}{2\lambda_t^2}\right)\right]$.
- Since this is a univariate density, one can compute its cdf by deterministic integration. One can then draw a random value by generating $u \sim U(0,1)$ and computing the $u\%$-quantile numerically, i.e. solving $u = CDF(\nu)$ for ν. The CDF of ν is different in each iteration of the Gibbs sampler since it depends on λ.
- Note that given λ, ν is independent of β and σ^2.

What prior for ν?

- It is difficult to discriminate between different large values of ν from the data. Intuition: whether $\nu = 30$ or 40, the Student density is close to the Gaussian.
- The posterior of ν when $\varphi(\nu) \propto 1$ has a long thick tail for 'large' values of ν (say $\nu > 20$). Actually, it does not tend to 0 as ν tends to ∞, hence is not integrable over $(0, \infty)$!
  Solutions:
  -truncate ν between e.g. $a = 0.05$ and $b = 30$;
  -use a prior that tends quickly enough to 0 when ν tends to ∞, so that the likelihood is dominated by the prior: $\varphi(\nu) \propto [1 + \nu^2]^{-1}$ if $\nu > 0$ (half Cauchy) is sufficient, but $\varphi(\nu) \propto 1/\nu$ is not, see Bauwens and Lubrano (1998, *Econometrics Journal*). An exponential density is another possible choice.

References

- For this chapter, see in BLR sections
  -2.7 and 5.3 about inference for regression under CP;
  -4.2 about elicitation of a prior;
  -4.3 about NIP;
  -4.5 about non-conjugate prior (Student-diffuse prior);
  -7.1 about heteroskedastic regression;
  -5.4 about autocorrelated errors.
- Bauwens L., Korobilis, D. (2011), Bayesian Methods, CORE DP 2011/61. Forthcoming in: N. Hashimzade and M. Thornton (Eds.), Handbook of Research Methods and Applications on Empirical Macroeconomics, Edward Elgaar Publishing.
  Covers linear regression with NIP, CP and non-CP.
- See list at the end of Chapter 1.

COURSE STRUCTURE

- Chapter 1: Concepts (p 2)
- Chapter 2: Numerical Methods (p 33)
- Chapter 3: Single Equation Regression Analysis (p 95)
- Chapter 4: VAR Models

CHAPTER 4

- 3.1 Unrestricted VAR and multivariate regression models

- 3.2 Posterior with NIP
- 3.3 Posterior with informative prior
- 3.4 The Minnesota prior
- 3.5 Restricted VAR and SURE models

VAR models

- We write a VAR model for the vector $x_t \in R^n$ as

$$A(L)x_t = c + \epsilon_t \tag{1}$$

  where $A(L) = I_n - A_1 L - A_2 L^2 - \cdots - A_p L^p$
  is a polynomial of degree p in the lag operator,
  A_i are square matrices ($n \times n$) of parameters,
  $c \in R^n$ is a vector of intercepts, and $\epsilon_t \sim I.N_n(0, \Sigma)$.
- No restrictions are imposed on the parameters c and A_i, implying that all equations of the system have the same explanatory variables (p lags of each variable in x_t).
- We can include other terms like a trend, dummy and other explanatory variables.

VAR as multivariate regression model

- The VAR system can be written as a multivariate regression model:

$$y_t = B' z_t + \epsilon_t, \qquad \epsilon_t \sim IN_n(0, \Sigma) \tag{2}$$

  where y_t, z_t, and B are of dimension $n \times 1$, $k \times 1$, and $k \times n$, respectively.

- The VAR model (1) corresponds to

$$\begin{aligned} y_t &= x_t \\ z_t &= (1 \; x'_{t-1} \; x'_{t-2} \; \dots \; x'_{t-p})' \\ B' &= (c \; A_1 \; A_2 \; \dots \; A_p) \\ k &= (n \times p) + 1. \end{aligned} \tag{3}$$

Matrix form of the model

- The matrix version of (2) for T observations (plus p initial ones in the VAR case) is obtained by transposing (2) and stacking:

$$Y = ZB + E, \qquad E \sim MN_{T\times n}(0, \Sigma \otimes I_T) \tag{4}$$

where

$$Y = \begin{bmatrix} y_1' \\ y_2' \\ \dots \\ y_T' \end{bmatrix}, \quad Z = \begin{bmatrix} z_1' \\ z_2' \\ \dots \\ z_T' \end{bmatrix}, \quad E = \begin{bmatrix} \epsilon_1' \\ \epsilon_2' \\ \dots \\ \epsilon_T' \end{bmatrix}.$$

- MN denotes a matricvariate normal distribution, which is a Normal distribution for a random matrix.

The matricvariate Normal distribution/1

- Let X denote a $p \times q$ random matrix and vec X its pq-dimensional column expansion.
- X is said to have a matricvariate Normal distribution with parameters $M \in \mathbb{R}^{p\times q}, P \in C_p$, and $Q \in C_q$, if and only if vec X has a multivariate Normal distribution with parameters vec M and $Q \otimes P$, i.e.

$$X \sim MN_{p\times q}(M, Q \otimes P) \Leftrightarrow \text{vec } X \sim N_{pq}(\text{vec } M, Q \otimes P).$$

- Therefore, its density function is given by

$$\begin{aligned} f_{MN}^{p\times q}(X|M, Q \otimes P) = {} & [(2\pi)^{pq}\, |P|^q\, |Q|^p]^{-1/2} \\ & \times \exp\{-\tfrac{1}{2}\text{tr}[Q^{-1}(X-M)'P^{-1}(X-M)]\}. \end{aligned} \tag{5}$$

The matricvariate Normal distribution/2

- The use of the trace operator in (5) originates from

$$
\begin{aligned}
&[\text{vec}\,(X-M)]'(Q\otimes P)^{-1}[\text{vec}\,(X-M)]\\
&=\Sigma_{i=1}^{q}\Sigma_{j=1}^{q}q^{ij}(x_i-m_i)'P^{-1}(x_j-m_j)\\
&=\text{tr}[Q^{-1}(X-M)'P^{-1}(X-M)],
\end{aligned}
$$

  where $x_i - m_i$ denotes the ith column of $X - M$ and and q^{ij} the (i,j)th element of Q^{-1}.

- All the properties of the multivariate Normal distribution apply to the matricvariate normal distribution through the vec operator. For details, see BLR, Appendix A.2.3.

The likelihood function

- Using (4) and applying (5):

$$\begin{aligned} L(B,\Sigma|d) &\propto |\Sigma|^{-T/2}\ \exp\{-\tfrac{1}{2}\text{tr}\,\Sigma^{-1}(Y-ZB)'(Y-ZB)\} \\ &= |\Sigma|^{-T/2}\ \exp\{-\tfrac{1}{2}\text{tr}\,\Sigma^{-1}[S+(B-\hat{B})'Z'Z(B-\hat{B})]\}. \end{aligned} \tag{6}$$

  where d stands for (Y,Z), and

$$\begin{aligned} \hat{B} &= (Z'Z)^{-1}Z'Y \\ S &= Y'M_ZY = Y'Y - Y'Z(Z'Z)^{-1}Z'Y. \end{aligned} \tag{7}$$

  We assume that $T > k+n+1$.

- As usual, we use kernels, i.e. we do not write the useless constants in the density.

CHAPTER 4

- 3.1 Unrestricted VAR and multivariate regression models
- 3.2 Posterior with NIP
- 3.3 Posterior with informative priors
- 3.4 The Minnesota prior
- 3.5 Restricted VAR and SURE models

Non-informative prior (NIP)

- It is usually defined as

$$\varphi(B, \Sigma) \propto |\Sigma|^{-(n+1)/2}. \tag{8}$$

- This can be interpreted as uniform on all the elements of B on $\mathbb{R}^{kn}$, and uniform on the elements of Σ taking into account that Σ is a covariance matrix (thus symmetric and positive-definite).
- We shall see that $|\Sigma|^{-(n+1)/2}$ can be seen as the limit of a proper distribution.
- The prior is 'improper': it does not integrate to 1. However this does not prevent the posterior to be integrable (or proper).

Posterior with NIP

- By multiplication of (8) and (6), we get the posterior kernel, and we can see that its corresponds to a MN for B given Σ and an inverted Wishart density for Σ:

$$\begin{aligned} \varphi(B,\Sigma|Y,Z) &\propto |\Sigma|^{-(T+n+1)/2} \\ &\quad \exp\{-\tfrac{1}{2}\text{tr}\,\Sigma^{-1}[S+(B-\hat{B})'Z'Z(B-\hat{B})]\} \\ &\propto \underbrace{f_{MN}^{k\times n}(B|\hat{B},\Sigma\otimes(Z'Z)^{-1})}_{\varphi(B|\Sigma,d)}\ \underbrace{f_{IW}^{n}(\Sigma|T-k,S)}_{\varphi(\Sigma|d)}. \end{aligned} \tag{9}$$

- $f_{IW}^{n}(\Sigma|T-k,S)$ denotes an inverted Wishart density for the matrix Σ, with parameters S and $T-k$. Hence

$$\mathsf{E}(\Sigma|d)=\frac{1}{T-k-n-1}S \quad \text{if } T>k+n+1.$$

The inverted Wishart distribution/1

- A random matrix $\Sigma \in C_q = \{\Sigma \mid \Sigma$ is $q \times q$ and PDS$\}$ has an inverted Wishart distribution with parameters $S \in C_q$ and $\nu > q-1$, i.e. $\Sigma \sim IW_q(\nu, S)$, if its density function is given by

$$f_{IW}^{q}(\Sigma|\nu, S) = C_{IW}^{-1}(\nu, S; q)\, |\Sigma|^{-\frac{1}{2}(\nu+q+1)} \exp\left[-\frac{1}{2}\text{tr}(\Sigma^{-1}S)\right], \tag{10}$$

where

$$C_{IW}(\nu, S; q) = 2^{\frac{1}{2}\nu q}\, \pi^{\frac{1}{4}q(q-1)} \prod_{i=1}^{q} \Gamma\left(\frac{\nu+1-i}{2}\right) |S|^{-\frac{1}{2}\nu}. \tag{11}$$

- For $q = 1$, Σ and S are scalar and the distribution is equivalent to the inverted-gamma-2 density.

The inverted Wishart distribution/2

- For $\nu > q + 1$, the expectation of Σ is given by

$$\mathsf{E}(\Sigma) = \frac{1}{\nu - q - 1} S. \tag{12}$$

- If Σ_{11} is extracted from $\Sigma \sim IW_q(\nu, S)$ along its first q_1 rows and columns, the marginal distribution of Σ_{11} is $IW_{q_1}(\nu - q_2, S_{11})$ where S_{11} is extracted from S along its first q_1 rows and columns and $q_2 = q - q_1$.
- For other properties of the inverted Wishart distribution, see BLR, Section A.2.6.
- If $\Sigma \sim IW_q(\nu, S)$, the distribution of Σ^{-1} is said to be a Wishart distribution with parameters ν and S..

Posterior with NIP

- The posterior density of B and Σ can also be factorized as

  $$\varphi(B,\Sigma|d) = \varphi(B|d)\varphi(\Sigma|B,d)$$

  where

  $$\begin{aligned} B|d &\sim Mt_{k\times n}(\hat{B}, Z'Z, S, T-k) \\ \Sigma|B,d &\sim IW_n(T,(Y-ZB)'(Y-ZB)) \end{aligned} \tag{13}$$

- The marginal distribution of B is a matricvariate-Student (or matricvariate-t). This implies that

  $$\begin{aligned} \text{E}(\text{vec } B|d) &= \text{vec } \hat{B} \quad \text{if } T > k+n, \\ \text{Var}(\text{vec } B|d) &= \frac{1}{T-k-n-1} S \otimes (Z'Z)^{-1} \quad \text{if } T > k+n+1. \end{aligned} \tag{14}$$

Direct sampling under NIP

For computing posterior features that are not known analytically (such as impulse responses, eigenvalues...), direct sampling of the posterior can be used to generate R draws of B and Σ:

1. Set $r = 1$.
2. *Based on based on (9) draw successively*
   $\Sigma^{(r)}$ *from* $IW_n(T-k, S)$ and
   $B^{(r)}$ *from* the $MN(\hat{B}, \Sigma^{(r)} \otimes (Z'Z)^{-1})$.
3. Set $r = r + 1$ *and go to step 2 unless* $r > R$.
4. Compute what you are interested in from the R *draws.*

Gibbs sampling for posterior under NIP

Another (less efficient) option is a Gibbs sampling algorithm to generate R draws from the posterior of B and Σ after R_0 warming-up draws:

1. *Choose an initial value* $\Sigma^{(0)}$ *(e.g.* $S/(T-k-n-1)$*).*
2. *Set* $r = 1$.
3. *Draw successively* $B^{(r)}$ *from* $MN_{k\times n}\left(\hat{B}, \Sigma^{(r-1)} \otimes (Z'Z)^{-1}\right)$ and $\Sigma^{(r)}$ *from* $IW_n(T, (Y - ZB^{(r)})'(Y - ZB^{(r)}))$.
4. *Set* $r = r + 1$ *and go to step 3 unless* $r > R_0 + R$.
5. *Discard the first* R_0 *values of* $B^{(r)}$ *and* $\Sigma^{(r)}$*. Compute what you are interested in from the last* R *draws.*

CHAPTER 4

- 3.1 Unrestricted VAR and multivariate regression models
- 3.2 Posterior with NIP
- 3.3 Posterior with informative priors
- 3.4 The Minnesota prior
- 3.5 Restricted VAR and SURE models

Conjugate prior

- The conjugate prior is a MN for $B|\Sigma$ times an IW for Σ:
  - $B|\Sigma \sim MN_{k\times n}(B_0, \Sigma \otimes M_0^{-1})$,
  - $\Sigma \sim IW_n(\nu_0, S_0)$.

- Thus, assuming $\nu_0 > n-1$, $\text{Var}(\text{vec } B) = \frac{1}{\nu_0-n-1} S_0 \otimes M_0^{-1}$. Hence, the prior covariance matrix of B_j, the coefficients of equation j (they are stacked in column j of B), is $\text{Var}(B_j) = \frac{1}{\nu_0-n-1} S_{0,jj} M_0^{-1}$.

- Thus, the prior covariance of the coefficients of two different equations are proportional to each other and have the same correlation structure.
  This is very restrictive, hence the CP is never used.

Extended conjugate prior

- An extended conjugate prior has been defined in the literature (see references). It avoids the restriction explained above. However, as the CP, it requires to be informative on Σ in order to be informative on B (as in the single equation case).
- It is much easier, for eliciting the prior on B, to remain non-informative (or diffuse) on Σ, with $\varphi(\Sigma) \propto |\Sigma|^{-(n+1)/2}$. Notice that this is the limit of the conjugate IW prior obtained by setting $S_0 = 0$ and $\nu_0 = 0$ (not $\nu_0 = n - 1$).
- Then one can use "any" desired (non-conjugate) prior on B (or some elements of B), and use the same Gibbs sampler as sketched for the NIP case, with a MH step to draw B if the conditional density of $B|\Sigma, d$ cannot be simulated directly.

Normal diffuse prior

- If the prior on vec B is Normal, then the conditional density of $B|\Sigma, d$ is Normal and is easy to simulate.
- Indeed, let vec $B \sim N_h($vec $B_0, M_0^{-1})$ where $h = k \times n$ is the dimension of vec B. Then, assuming the diffuse prior for Σ, the posterior density of B and Σ is proportional to

$$
\begin{aligned}
&|\Sigma|^{-(T+n+1)/2} \exp\{-\tfrac{1}{2}\text{tr}(\Sigma^{-1}S)\} \\
&\exp\{-\tfrac{1}{2}(\text{vec } B - \text{vec } \hat{B})'(\Sigma^{-1} \otimes Z'Z)(\text{vec } B - \text{vec } \hat{B})\} \\
&\exp\{-\tfrac{1}{2}(\text{vec } B - \text{vec } B_0)' M_0 (\text{vec } B - \text{vec } B_0)\}
\end{aligned}
\tag{15}
$$

- By adding the arguments of the two $\exp$ functions and combining the two quadratic forms in vec B into a single quadratic form, one can show that vec $B|\Sigma, d$ is Normal.

Density of $B|\Sigma, d$

- The result is

$$\text{vec } B|\Sigma, d \sim N_h(\text{vec } B_*, M_*^{-1}) \qquad (16)$$

where

$$\begin{aligned} M_* &= \Sigma^{-1} \otimes Z'Z + M_0 \\ \text{vec } B^* &= M_*^{-1} \left[(\Sigma^{-1} \otimes Z'Z)\text{vec } \hat{B} + M_0 \text{vec } B_0 \right] \end{aligned} \qquad (17)$$

- The mean and the covariance matrix of vec B depend on Σ through the likelihood contribution.
- Notice that if $M_0 = 0$, the prior on B is non-informative (then B_0 can be set to 0), and then $B|\Sigma, d$ is the same as in (9).

Density of $\Sigma|B,d$

- In (15), the first three factors only depend on Σ, hence they correspond to the kernel of $\Sigma|B,d$. The kernel of the posterior can be expressed as

$$|\Sigma|^{-(T+n+1)/2}\exp\{-\frac{1}{2}\text{tr}\left[\Sigma^{-1}(Y-ZB)'(Y-ZB)\right]\}.$$

- Hence, it is clear that

$$\Sigma|B,d\sim IW_n(T,(Y-ZB)'(Y-ZB)). \qquad (18)$$

This is the same as in (14).

Gibbs sampling under Normal-diffuse prior

For computing posterior features, here is a Gibbs sampling algorithm to generate R draws from the posterior of B and Σ (after R_0 warming-up draws):

1. *Choose an initial value* $\Sigma^{(0)}$ *(e.g.* $S/(T-k-n-1)$*).*
2. *Set* $r = 1$.
3. *Draw successively* vec $B^{(r)}$ *from* the Normal density in (16) where vec B_* and M_* are computed with $\Sigma = \Sigma^{(r-1)}$, and $\Sigma^{(r)}$ *from* $IW_n(T, (Y - ZB^{(r)})'(Y - ZB^{(r)}))$.
4. *Set* $r = r + 1$ *and go to step 3 unless* $r > R_0 + R$.
5. *Discard the first* R_0 *values of* $B^{(r)}$ *and* $\Sigma^{(r)}$*. Compute what you are interested in from the last* R *draws.*

Non-Normal prior on vec B

- If a non-Normal prior $\varphi(\text{vec } B)$ is used, the conditional posterior $B|\Sigma, d$ is proportional to

$$|\Sigma|^{-(T+n+1)/2} \;\; \varphi(\text{vec } B)$$
$$\exp\{-\tfrac{1}{2}(\text{vec } B - \text{vec } \hat{B})'(\Sigma^{-1} \otimes Z'Z)(\text{vec } B - \text{vec } \hat{B})\}$$

  and not a known density that can be simulated directly.

- The previous Gibbs algorithm must be adapted, by simulating vec $B|\Sigma = \Sigma^{(r-1)}$ using a MH step.
- The proposal density can be constructed by approximating $\varphi(\text{vec } B)$ by a Normal density, and using (16) as proposal. If $\varphi(\text{vec } B)$ is far from being Normal, a better proposal should be designed (e.g. a mixture of Normal densities).

CHAPTER 4

- 3.1 Unrestricted VAR and multivariate regression models
- 3.2 Posterior with NIP
- 3.3 Posterior with informative priors
- 3.4 The Minnesota prior
- 3.5 Restricted VAR and SURE models

Origin

- Litterman and Sims have defined a prior for the VAR model (1), see e.g. Doan, Litterman, and Sims (1984) for details.
- It is called the ‘Minnesota’ (or ‘Litterman’) prior in the literature since Litterman wrote his doctoral dissertation at the University of Minnesota.
- This prior is informative on all the coefficients of the A_i matrices, and non-informative on the other parameters.

Prior expectation

- If x_t is a set of series in levels, the prior expectation is that the VAR system consists of n random walks, i.e. the prior mean of A_i is zero for $i \geq 2$, and the prior mean of A_1 is equal to I_n (the identity matrix).
- For quarterly data with a seasonal pattern, it is the prior mean of A_4 that should be an identity matrix.
- If x_t is the first difference of a set of series, there should be no identity matrix in the prior mean, only zero prior means.

Prior covariance matrix/1

- The prior covariance matrix of all the parameters in the A_i matrices is diagonal.
- For a given equation of the VAR, the standard deviation of the corresponding diagonal element of A_1 is a fixed value (say λ), meaning that one is of course not sure that this parameter is equal to one.
- The standard deviation of the coefficient of lag i of the same variable is equal to λ/i, reflecting the idea that the larger the lag, the more likely the coefficient is to be close to zero.

Prior covariance matrix/2

- The standard deviations of the coefficients of the lags of every other variable in the equation have the same decreasing pattern.
- For lag i of the variable x_j in equation k, the standard deviation is $\lambda\theta\sigma_k/i\sigma_j$, where θ is a scalar between 0 and 1 to incorporate the idea that the lags of x_j $(j \neq k)$ are more likely to have zero coefficients than the lags of x_k in equation k.
- The ratio σ_k/σ_j of the standard deviations of the error terms is a way to take account of the difference in the scale of the different variables.
- For a given equation, the choice of the prior moments requires two values, λ and θ, which can be the same for all the equations.

Example

- For the first equation of a bivariate VAR with two lags, the prior means and standard deviations are given in parentheses:

$$\begin{aligned}
x_{1,t} = \quad & \alpha_{11} x_{1,t-1} && + \quad \alpha_{12} x_{1,t-2} \\
& (1, \lambda) && \quad\;\; (0, \lambda/2) \\
& + \beta_{11} x_{2,t-1} && + \quad \beta_{12} x_{2,t-2} && + \epsilon_{1,t}. \\
& (0, \theta\lambda\sigma_1/\sigma_2) && \quad\;\; (0, \theta\lambda\sigma_1/2\sigma_2)
\end{aligned}$$

Minnesota prior as Normal prior/1

- The Minnesota prior is a Normal distribution with the described mean and diagonal covariance matrix:

$$\text{vec } B|\sigma_1, \sigma_2, \ldots, \sigma_n \sim N_h(\underbrace{\text{vec } B_0}_{\beta_0}, M_0^{-1}) \qquad (19)$$

  where σ_i $(i = 1 \text{ to } n)$ is the square root of the ith diagonal element of the covariance matrix Σ.

- Note that M_0^{-1} depends on these parameters. To make this prior easy to specify, and the posterior sampling easy, one can set σ_i equal to the ML estimate of the i-th equation of the VAR. Then, one can use the Gibbs sampling algorithm sketched for the Normal-diffuse prior.

Minnesota prior as Normal prior/2

- The non-zero diagonal elements of M_0 are the inverses of the variances of the coefficients on which one is informative (as described above). Only two scalars have to be chosen: λ and θ.
- The zero diagonal elements of M_0 correspond to the parameters on which one is not informative (like the intercepts), and the corresponding elements of β_0 are set to zero.
- So β_0 consists of ones for the diagonal elements of A_1 and zeros everywhere else.

Impact of the Minnesota prior

- The influence of the Minnesota prior on the posterior results of the VAR coefficients is twofold:
- 1) The precision of the 'estimates' is improved because of the adding up of prior and sample precisions.
- 2) The posterior means of the coefficients on which the sample is weakly informative are shrunk towards the prior means (most of them being null), and away from the least squares (LS) estimates (which are the posterior means under the diffuse prior).
- In VAR models, it is customary to find LS estimates which are very imprecisely determined, so the prior may help to shrink these coefficients to less 'extreme' values than the LS values.

Impact of the Minnesota prior

- This usually helps to improve the predictions of the model. For a set of macroeconomic series, the VAR model with the Minnesota prior has indeed been found to be often a better prediction tool than the VAR model without the prior (i.e. least squares predictions); see Litterman (1986) for an account of such comparisons. Anoher interesting paper in that respect is that of Kadiyala and Karlsson (1997).

CHAPTER 4

- 3.1 Unrestricted VAR and multivariate regression models
- 3.2 Posterior with NIP
- 3.3 Posterior with informative priors
- 3.4 The Minnesota prior
- 3.5 Restricted VAR and SURE models

Exclusion restrictions in VAR

- One may wish to impose that the explanatory variables are not the same in all equations. Examples are
  -Granger non-causality restrictions (lags of one variable excluded in some equations);
  -Inclusion of a linear trend or seasonal dummy variables in some equations but not in others.
- Such exclusion restrictions complicate the derivation of the posterior results even under the diffuse prior (8), where B is replaced by B_c (the constrained B):

$$\varphi(B_c, \Sigma) \propto |\Sigma|^{-(n+1)/2}. \tag{20}$$

Posterior under NIP in restricted VAR

- The posterior density is still like in the first line of (9):

$$\varphi(B_c, \Sigma | Y, Z) \propto |\Sigma|^{-(T+n+1)/2} \exp\{-\tfrac{1}{2}\text{tr}\,\Sigma^{-1}[S + (B_c - \hat{B})' Z' Z (B_c - \hat{B})]\} \tag{21}$$

- The conditional posterior of $B_c|\Sigma$ implied by (21) is not in general a MN distribution with expectation $\hat{B}$ and covariance matrix $\Sigma \otimes (Z'Z)^{-1}$ as in the second line of (9).
- We know the posterior kernel of $B_c|\Sigma$ but we cannot simulate it directly. For example, rejection sampling from the unconstrained MN in (9) is not feasible: with exact restrictions, all draws will be rejected.

Posterior under NIP in restricted VAR

- However, the conditional posterior of $\Sigma|B_c$ remains an inverted Wishart density.
- Neither direct sampling nor Gibbs sampling as defined in Section 3.1 can be used.
- One solution is to replace the sampling of $B|\Sigma$ as MN in the unconstrained model by a MH step for $B_c|\Sigma$ in the constrained model. This requires to design a good enough proposal density.
- Another approach is based on writing the VAR subject to exclusion restrictions as a system of seemingly unrelated regression equations (SURE).

SURE system/1

- A SURE system is a set of regression equations, possibly with different regressors, whose error terms are correlated. It can be written as

$$Y_i = Z_i\beta_i + E_i, \qquad i = 1, \ldots, n \tag{22}$$

  where Y_i, Z_i, and β_i are of dimension $T \times 1$, $T \times k_i$, and $k_i \times 1$ respectively.

- In compact matrix format, we write

$$y = \mathcal{Z}\beta + \epsilon \tag{23}$$

- The distribution of the $Tn \times 1$ vector ϵ is assumed to be $N_{Tn}(0, \Sigma \otimes I_T)$; this is actually the same hypothesis as in (4) since $\epsilon =$ vec E.

SURE system/2

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \qquad \epsilon = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{bmatrix} \tag{24}$$

and

$$\mathcal{Z} = \begin{bmatrix} Z_1 & 0 & \ldots & 0 \\ 0 & Z_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \vdots & Z_n \end{bmatrix} \tag{25}$$

SURE system/3

Another useful writing of the SURE system (23) is

$$Y = WB_c + E \tag{26}$$

with Y and E as in (4), $W = (Z_1 \; Z_2 \; \ldots \; Z_n)$ and

$$B_c = \begin{bmatrix} \beta_1 & 0 & \ldots & 0 \\ 0 & \beta_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \vdots & \beta_n \end{bmatrix} \tag{27}$$

The matrix W is not of full column rank if some equations share the same explanatory variables (e.g. a constant).

Constrained VAR as SURE

An example of a bivariate VAR with one lag and non-causality of x_2 for x_1 that can be put easily in the form of (23) and (26) is:

$$\begin{aligned} x_{1,t} &= \beta_{11} x_{1,t-1} + c_1 + \epsilon_{1,t} \\ x_{2,t} &= \beta_{21} x_{1,t-1} + \beta_{22} x_{2,t-1} + c_2 + \epsilon_{2,t}. \end{aligned}$$

Thus
$\beta_1 = (\beta_{11}\ c_1)'$,
$\beta_2 = (\beta_{21}\ \beta_{22}\ c_2)'$,
Z_1 has $(x_{1,t-1}\ 1)$ as its t-th row, and
Z_2 has $(x_{1,t-1}\ x_{2,t-1}\ 1)$ as its t-th row.
So $W = (Z_1\ Z_2)$ has rank 3.

Posterior density of SURE with NIP

- Posterior marginal densities for the SURE system are not available analytically, but the conditional posterior densities of $\beta|\Sigma$ and $\Sigma|\beta$ are known.
- Hence they can be used to define a Gibbs sampling algorithm cycling between these two densities. See Percy, D.F. (1992).

Conditional posterior densities

With the diffuse prior (20), which can be written as $\varphi(\beta, \Sigma) \propto |\Sigma|^{-(n+1)/2}$ since β includes the unconstrained parameters in B_c, the conditional posterior densities are:

$$\begin{aligned} \beta|\Sigma, d &\sim N_K(\hat{\beta}, \left[\mathcal{Z}'(\Sigma^{-1} \otimes I_T)\mathcal{Z}\right]^{-1}) \\ \Sigma|\beta, d &\sim IW_n(T, Q) \end{aligned} \tag{28}$$

where $K = \sum_{i=1}^{n} k_i$ and

$$\begin{aligned} \hat{\beta} &= \left[\mathcal{Z}'(\Sigma^{-1} \otimes I_T)\mathcal{Z}\right]^{-1} \mathcal{Z}'(\Sigma^{-1} \otimes I_T)y \\ Q &= (Y - WB_c)'(Y - WB_c). \end{aligned} \tag{29}$$

NB: If $Z_1 = Z_2 = \cdots = Z_n = Z$ of (4) in the SURE formulation (22), $\hat{\beta}$ of (29) is equal to vec $\hat{B}$ of (7).

Proof of (28)

- (23) is a linear regression model, so that

$$\varphi(\beta,\Sigma|d) \propto$$
$$|\Sigma|^{-(T+n+1)/2} \exp\left[-\tfrac{1}{2}\,(y-\mathcal{Z}\beta)'(\Sigma^{-1}\otimes I_T)(y-\mathcal{Z}\beta)\right]$$
$$= |\Sigma|^{-(T+n+1)/2} \exp -\tfrac{1}{2}\left\{\left[s+(\beta-\hat\beta)'\mathcal{Z}'(\Sigma^{-1}\otimes I_T)\mathcal{Z}(\beta-\hat\beta)\right]\right\}$$

where $s = y'(\Sigma^{-1}\otimes I_T)y - \hat\beta'\mathcal{Z}'(\Sigma^{-1}\otimes I_T)\mathcal{Z}\hat\beta$.

The posterior conditional density of $\beta|\Sigma$ follows directly.

- To obtain the density of $\Sigma|\beta$, we use (26), so that

$$\varphi(\beta,\Sigma|d) \propto |\Sigma|^{-(T+n+1)/2}\ \exp\left[-\tfrac{1}{2}\,\mathrm{tr}\,\Sigma^{-1}(Y-WB_c)'(Y-WB_c)\right],$$

which is recognized as the kernel of an IW density.

References

- For this chapter, see BLR section 9.2.. BLR, section 9.3 covers Bayesian inference for the VAR model with cointegration relations, also called the vector error correction model (VECM).
- About Bayesian inference for SURE systems:
  Percy, D.F. (1992), Prediction for seemingly unrelated regressions, JRSS B, 54, 243-252.
- For the extended conjugate prior, see
  Drèze, J.H. and Richard, J.-F. (19833), Bayesian analysis of simultaneous equation systems, Chapter 9 of the *Handbook of Econometrics*, Volume I, edited by Griliches Z. and Intriligator M.D.,
  and the references cited in that chapter.

References

About the Minnesota prior:

- Doan, Litterman, and Sims (1984), Forecasting and conditional projection under realistic prior distributions. *Econometric Reviews* 3, 1-100.
- Litterman, R.B. (1986), Forecasting with Bayesian vector autoregressions, *Journal of Business and Economic Statistics*, 4, 25-38.
- Kadiyala and Karlsson (1997), Numerical ùethodsfor estimation and inference in Bayesian VAR models, *Journal of Applied Econometrics* 12, 99-132.

References

- See also the reference list at the end of Chapter 1 of this course. In particular the chapter on Macroeconomic Applications in the Oxford Handbook of Bayesian Econometrics covers the above topics and other such as DSGE models and time-varying parameter VAR models.
- Koop, G. and Korobilis, D. (2010), Bayesian Multivariate Time Series Methods for Empirical Macroeconomics. Foundations and Trends in Econometrics, Vol.3, No.4, 267-358.
  A discussion paper version is available at http://personal.strath.ac.uk/gary.koop/kk3.pdf, and Matlab codes at https://sites.google.com/site/dimitriskorobilis/matlab/code-for-vars