Relatively informative priors and model selection in latent largely Gaussian models

Øyvind Hoveid

Norwegian Institute of Bioeconomic Research,
Mail: oyvind.hoveid@nibio.no

AMLGM, Trondheim, September 17-18, 2015

Introduction

Gaussian distributed latent variables are unsurpassed in hierarchical/spatio/temporal models due to the equivalence of contingent independence and zero cross-precision (Rue and Held, 2005). However, many real world (e.g. economic) phenomena require other distributions. I will here point to a broader class of models in which the Gaussian latents depend on non-Gaussian shocks, and the predictors of observations are transformed Gaussian. Within the class it is necessary to keep determinants of Gaussian densities non-singular. Informative Inverted Wishart priors on variance matrices guard effectively against this situation. The idea stems from the term "Boundary avoiding priors" introduced by Gelman (2014). The priors are changed from Gamma to Inverse Wishart, though.

Informative priors need a justification. I will promote them as acceptable with a scaling factor $\omega \in (\mathbf{0}, \mathbf{1})$ for error variance determinants relative to the sum of latent and error variance determinants. With $\omega = \mathbf{0}$, error variance is singular. With $\omega = \mathbf{1}$, latent variance is singular. Estimation will take various fixed interior ω. A final value will be found according to a model selection criterion that rates models according to likelihood less a penalty for model complexity. Smaller ω means more complexity. Watanabe's information criterion (WAIC) (Watanabe 2010, 2013), (Gelman, Hwang, Vehtari 2013) works well for such models. With the framework on top, the priors are denoted *relatively informative priors* (RIP).

Model class

Consider models of the following class:

$$\phi(\boldsymbol{y} - \boldsymbol{Bx}, \boldsymbol{R}),\ \boldsymbol{x} \equiv \boldsymbol{F}(\boldsymbol{w}),\ \phi(\boldsymbol{w} - \boldsymbol{Av}, \boldsymbol{Q}),\ \pi(\boldsymbol{v}|\boldsymbol{P})$$
$$\mathcal{W}^{-1}(\boldsymbol{P}, \boldsymbol{Q}),\ \mathcal{W}^{-1}(\boldsymbol{R}|\omega, |\boldsymbol{P}|, |\boldsymbol{Q}|),\ \omega \in (\mathbf{0}, \mathbf{1})$$

where

- $\boldsymbol{y}$ is observations with multivariate Gaussian distribution ϕ
- $\boldsymbol{Bx}$ is the predictor of $\boldsymbol{y}$ selected from latents $\boldsymbol{x}$ with selector matrix $\boldsymbol{B}$ (with one non-zero element equal to $\mathbf{1}$ in each row)
- $\boldsymbol{R}$ is noise variance
- $\boldsymbol{F}$ is a 1-1 mapping from Gaussian latents $\boldsymbol{w}$ to predictors $\boldsymbol{x}$
- $\boldsymbol{Av}$ is the predictor of $\boldsymbol{w}$ generated from the vector $\boldsymbol{v}$ with a selector matrix $\boldsymbol{A}$
- $\boldsymbol{Q}$ is the variance matrix of $\boldsymbol{w}$, with a dependence structure in terms of a sparse $\boldsymbol{Q}^{-1}$. $\boldsymbol{w}$ is sorted to orient dependencies towards NW. Sparsity is then preserved in Cholesky factors with reversed Cholesky factorization: $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^{-\frac{T}{2}}\boldsymbol{Q}^{-\frac{1}{2}}$
- $\boldsymbol{v}$ is a vector of independent non-Gaussian variables with a diagonal scale matrix $\boldsymbol{P}$

Model selection within the fully Gaussian frame

Consider the total likelihood:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{v}) = \phi(\boldsymbol{y} - \boldsymbol{Bx}, \boldsymbol{R})\phi(\boldsymbol{x} - \boldsymbol{Av}, \boldsymbol{Q})$$

with posterior predictor of $\boldsymbol{x}$:

$$\widehat{\boldsymbol{x}}(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{y}) = \boldsymbol{H}\left(\boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{y} + \boldsymbol{Q}^{-1}\boldsymbol{Av}\right)$$

where

$$\boldsymbol{H} = (\boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B} + \boldsymbol{Q}^{-1})^{-1}$$

With flat priors, the hyper-parameters $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{v})$, can be varied until maximum likelihood. The likelihood can be stated as the total Gaussian one divided by the Laplace density of $\boldsymbol{x}$ at $\widehat{\boldsymbol{x}}$ (INLA). Analytically:

$$\begin{aligned}\log \pi(\boldsymbol{y}|\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{v}) &= \log\left[\phi(\boldsymbol{y} - \boldsymbol{B}\widehat{\boldsymbol{x}}, \boldsymbol{R})\phi\left(\widehat{\boldsymbol{x}} - \boldsymbol{Av}, \boldsymbol{Q}\right)/\phi(\mathbf{0}, \boldsymbol{H})\right] \\ &= \log\phi(\boldsymbol{y} + \boldsymbol{CAv}, \boldsymbol{G}) + \frac{1}{2}\left[(\boldsymbol{Av})^T\left(\boldsymbol{C}^T\boldsymbol{G}^{-1}\boldsymbol{C} - \boldsymbol{B}^T\boldsymbol{G}^{-1}\boldsymbol{B}\right)\boldsymbol{Av}\right]\end{aligned}$$

where

$$\boldsymbol{G} = \boldsymbol{R} + \boldsymbol{BQB}^T, \quad \boldsymbol{C} = \boldsymbol{GR}^{-1}\boldsymbol{BH}^{-1}\boldsymbol{Q}^{-1}$$

There is no chance of hitting $|\boldsymbol{G}| = \mathbf{0}$ in this case and no informative prior is needed, but the model selection issue remains. The constraint:

$$\omega|\boldsymbol{Q}| = (\mathbf{1} - \omega)|\boldsymbol{R}| \tag{1}$$

with models solved for different ω, will span out a set from which model selection can take place.

Model selection with transformed Gaussian latents and flat priors

Transformed Gaussian latents, $\boldsymbol{x} = \boldsymbol{F}(\boldsymbol{w})$, brings in a Jacobian determinant to state the prior distribution of $\boldsymbol{x}$:

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{v}) = \phi(\boldsymbol{y} - \boldsymbol{BF}(\boldsymbol{w}), \boldsymbol{R})\phi(\boldsymbol{w} - \boldsymbol{Av}, \boldsymbol{Q}))/|\partial\boldsymbol{F}(\boldsymbol{w})|$$

Some structure, say

$$\boldsymbol{F}(\boldsymbol{w}) = (\boldsymbol{F}_1(\boldsymbol{w}_1), \ldots, \boldsymbol{F}_t(\boldsymbol{w}_{1:t}))$$

will be helpful for the specification of the determinant by means of the Cholesky factor. Structures of contingent independence within $\boldsymbol{w}$ will only be disturbed as far as $\boldsymbol{F}_k(\boldsymbol{w})$ depends on $\boldsymbol{w}_j$ with $\boldsymbol{j} < \boldsymbol{k}$ while $\boldsymbol{Q}^{-1}_{kj} = \mathbf{0}$.

One might integrate $\mathcal{L}(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{v})$ with respect to $\boldsymbol{w}$ numerically, and search for a model of $\mathcal{L}(\boldsymbol{y}|\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{v})$. For sake of model selection one need not involve any constant ω, as $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{v})$ are constants themselves.

Model selection: transformed Gaussian latents and RIPs

Let $\boldsymbol{Q}$ and $\boldsymbol{R}$ be stochastic and introduce RIP-s, $\mathcal{W}^{-1}(\boldsymbol{Q}|\omega\mathcal{Q}, \boldsymbol{q})$ and $\mathcal{W}^{-1}(\boldsymbol{R}|(\mathbf{1} - \omega)\mathcal{R}, \boldsymbol{r})$. The constraint on determinants has now been softened with scaling of Wishart parameters $\omega\mathcal{Q}$ and $(\mathbf{1} - \omega)\mathcal{R}$ respectively. $\boldsymbol{q}$ and $\boldsymbol{r}$ are related to dimensionality and degrees of freedom. $\mathcal{Q}$ and $\mathcal{R}$ are fixed matrices proportionate to $\mathbf{Ex}\ \boldsymbol{Q}$ and $\mathbf{Ex}\ \boldsymbol{R}$ respectively.

Total likelihood is:

$$\mathcal{L}(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{w}, \boldsymbol{y}) = \phi(\boldsymbol{y} - \boldsymbol{BF}(\boldsymbol{w}), \boldsymbol{R})\phi(\boldsymbol{w} - \boldsymbol{Av}, \boldsymbol{Q})\mathcal{W}^{-1}(\boldsymbol{Q}|\omega\mathcal{Q}, \boldsymbol{q})\mathcal{W}^{-1}(\boldsymbol{R}|(\mathbf{1} - \omega)\mathcal{R}, \boldsymbol{r})/|\partial\boldsymbol{F}(\boldsymbol{w})|$$

and ignoring constants:

$$\begin{aligned}\log\mathcal{L}(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{w}, \boldsymbol{y}) &= \boldsymbol{q}\log|\boldsymbol{Q}^{-\frac{1}{2}}| + \boldsymbol{q}\log|\boldsymbol{R}^{-\frac{1}{2}}| - \log|\partial\boldsymbol{F}(\boldsymbol{w})| \\ &\quad - \frac{1}{2}\left[\mathbf{tr}\ \left((\boldsymbol{y} - \boldsymbol{BF}(\boldsymbol{w}))(\boldsymbol{y} - \boldsymbol{BF}(\boldsymbol{w}))^T + (\mathbf{1} - \omega)\mathcal{R}\right)\boldsymbol{R}^{-1} + \left((\boldsymbol{w} - \boldsymbol{Av})(\boldsymbol{w} - \boldsymbol{Av})^T + \omega\mathcal{Q}\right)\right.\end{aligned}$$

In terms of *all parameters* $\boldsymbol{u} = (\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{w})$, $\mathcal{L}(\boldsymbol{u}, \boldsymbol{y})$ is bounded and one can find the mode $\widehat{\boldsymbol{u}}$. The optimizing routine will deliver the negative Hessian matrix $\boldsymbol{H}^{-1}$ at the mode, from which the Laplace approximation is: $\pi(\boldsymbol{u}) = \phi(\boldsymbol{u} - \widehat{\boldsymbol{u}}, \boldsymbol{H})$. The Hessian matrix is expected to have a sparse factorization, $\boldsymbol{H}^{-1} = \boldsymbol{H}^{-\frac{T}{2}}\boldsymbol{H}^{-\frac{1}{2}}$. In order to sample from the Laplace distribution take $\boldsymbol{s} \in (\mathbf{0}, \mathbf{1})^{\boldsymbol{h}}$, with $\boldsymbol{h}$ the dimension of $\boldsymbol{H}$. Denoting $\boldsymbol{\Phi}$ as the cumulative standard normal distribution, $\boldsymbol{\Phi}^{-1}(\boldsymbol{s})$ is then a $\boldsymbol{h}$-dimensional standard normal, and the equation

$$\boldsymbol{H}^{-\frac{1}{2}}(\boldsymbol{u} - \widehat{\boldsymbol{u}}) = \boldsymbol{\Phi}^{-1}(\boldsymbol{s})$$

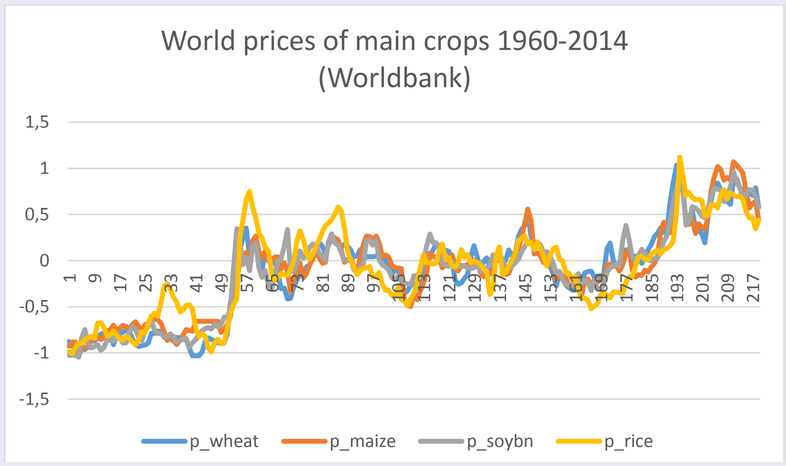
determines a sampled point. The log-probability of that selection is simply:

$$\log\pi(\boldsymbol{u}) = \log|\boldsymbol{H}^{-\frac{1}{2}}| - \frac{1}{2}(\boldsymbol{\Phi}^{-1}(\boldsymbol{s}))^2$$

A weighted sample from the posterior is obtained by using likelihoods divided by selection probabilities as weights (importance sampling).

An economic model

Modelling starts out from observations of food prices in the world market. Logged and scaled prices of wheat, maize, rice and soybeans are displayed below.



Latent variables provide an opportunity to recover some of the underlying economic mechanisms. Food prices are assumed determined by temporary equilibria where demand equals supply. Demand behaviour of a representative consumer is specified in terms of a cost function, $\boldsymbol{C}(\boldsymbol{p}_t, \boldsymbol{u}_t)$, specifying the minimum income to obtain utility level $\boldsymbol{u}_t$ at market prices $\boldsymbol{p}_t$. The structure of cost minimization then leads to the equations,

$$\boldsymbol{c}_t = \partial_{\boldsymbol{p}}\boldsymbol{C}(\boldsymbol{p}_t, \boldsymbol{u}_t)$$
$$\boldsymbol{i}_t = \boldsymbol{C}(\boldsymbol{p}_t, \boldsymbol{u}_t)$$

where $\boldsymbol{c}_t$ is consumed quantity and $\boldsymbol{i}_t$ is income. The functional specification of $\boldsymbol{C}$ is a CES (constant elasticity of substitution) function:

$$\boldsymbol{C}(\boldsymbol{p}, \boldsymbol{u}) = \left[\sum_{\boldsymbol{k}}(\beta_{\boldsymbol{k}}(\boldsymbol{u})\boldsymbol{p}_{\boldsymbol{k}})^{\rho}\right]^{\frac{1}{\rho}}, \quad \partial_{\boldsymbol{p}}\boldsymbol{C}(\boldsymbol{p}, \boldsymbol{u}) = \beta(\boldsymbol{u})^T \cdot \left[\frac{\beta(\boldsymbol{u})^T \cdot \boldsymbol{p}^T}{\boldsymbol{C}(\boldsymbol{p}, \boldsymbol{u})}\right]^{\rho-1}$$

With respect to production: it takes a long time to produce food, even longer times to increase capacity in periods of high prices, and the outcome is always uncertain. The production is consequently modelled as a long-lagged temporary process in which nature provide independent shocks, $\boldsymbol{v}_t$.

Translation to the statistical model:

- Observations $\boldsymbol{y}_{\boldsymbol{tk}}$: logged prices and income at $\boldsymbol{t}$, $= \log \boldsymbol{p}_{\boldsymbol{tk}}$ with $\boldsymbol{k} \in \{\text{wheat, sugar, soybeans, maize, other}\}$, and also $\boldsymbol{y}_{\boldsymbol{t},'\boldsymbol{i}'} = \log \boldsymbol{i}_{\boldsymbol{t}}$
- Predictors $\boldsymbol{x}_{\boldsymbol{tk}}$: of all observations $\boldsymbol{y}_{\boldsymbol{tk}}$, and also, $\boldsymbol{x}_{\boldsymbol{t},'\boldsymbol{u}'} = \boldsymbol{u}_{\boldsymbol{t}}$
- Gaussian variables $\boldsymbol{w}_{\boldsymbol{tk}} = \log \boldsymbol{c}_{\boldsymbol{tk}}$ with temporal structure, $\boldsymbol{Q}$ and expectation $\boldsymbol{v}_{\boldsymbol{tk}}$, and also independent parameters: $(\boldsymbol{w}_{0\boldsymbol{k}} = \log\beta_{\boldsymbol{k}}, \boldsymbol{w}_{00} = \log\rho)$
- Transformation $\boldsymbol{x} = \boldsymbol{F}(\boldsymbol{w})$ defined implicitly by $\exp(\boldsymbol{w}_t) = \partial_{\boldsymbol{p}}\boldsymbol{C}(\boldsymbol{x}_t)$ and $\exp(\boldsymbol{x}_{\boldsymbol{t},'\boldsymbol{i}'}) = \boldsymbol{C}(\boldsymbol{x}_t)$
- Independent Cauchy shocks: $\boldsymbol{v}_{\boldsymbol{tk}}$ for each crop $\boldsymbol{k}$

Relevant issues:

- Which time dependencies dominate $\boldsymbol{Q}^{-1}$?
- What is the better structure of the cost function, $\boldsymbol{C}(\boldsymbol{p}, \boldsymbol{u})$?
- What is the risk of the consumer due to food price fluctuations?

NIBIO
NORSK INSTITUTT FOR BIOØKONOMI