Statistica Sinica Preprint No: SS-2024-0052	
Title	Bayesian Statistics by Arithmetic Operations of Conjugate Distributions
Manuscript ID	SS-2024-0052
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0052
Complete List of Authors	Hang Qian
Corresponding Authors	Hang Qian
E-mails	matlabist@gmail.com
Notice: Accepted version subject to English editing.	

Bayesian Statistics by Arithmetic Operations of Conjugate Distributions

Hang Qian

The MathWorks, Inc.

Abstract: Conjugate distributions provide an entry point to Bayesian analysis. By defining summation, subtraction, and multiplication operators for conjugate distributions, we study Bayesian statistics by arithmetic operations. A striking feature is that the non-informative prior fulfills the central role of zero in mathematics. The summation operator connects Bayesian and frequentist estimators by a simple equation, which also provides an efficient method for evaluating the marginal likelihood. The subtraction operator facilitates cross-validation, rolling-window estimation, and regression under multicollinearity. The multiplication operator simplifies the weighted regression with a discount factor. Arithmetic operations conceptualize pseudo data in the conjugate prior, sufficient statistics that determine the likelihood, and the posterior that balances the prior and data.

Key words and phrases: Conjugacy, Exponential family, Linear regression, Statistics education

1. Introduction

On the one hand, summation, subtraction, and multiplication are basic arithmetic operators essential for primary education of mathematics. On the other hand, normal, beta and gamma priors are common conjugate distributions essential for introductory courses of Bayesian statistics. Traditionally, they are separate topics studied in different areas. We bridge the gap by defining arithmetic operators for conjugate distributions. Bayesian textbook materials, such as normal linear regressions, beta-binomial and gamma-Poisson models, are presented from a new perspective. The conjugate arithmetic operations provide novel pedagogical methods for studying 1) the conjugate prior that incorporates pseudo data, 2) the non-informative prior that accomplishes the role of zero in mathematics, 3) the sufficient statistics that determine the likelihood function, and 4) the posterior that balances data and the prior information.

Section 2 is the core of the paper, focusing on the theory and applications of normal-inverse-gamma (NIG) arithmetic for linear regressions. Section 3 extends the summation operator to distributions in the exponential family. Section 4 provides an application of rolling-window regressions. Section 5 concludes the paper.

2. Normal Linear Regressions

Linear regressions are presented in most textbooks of Bayesian statistics and econometrics. See Gelman et al. (2014, p.353), Bolstad and Curran (2017, p.411), Christensen et al. (2011, p.223), Koop (2003, p.33), among others. A normal linear regression is specified as

$$Y = X\beta + \sigma\varepsilon, \tag{2.1}$$

where Y and X are $n \times 1$ and $n \times d$ matrices for the response variable and predictors. The disturbances ε follow the standard normal distributions.

2.1 Priors and Posteriors

The normal likelihood is derived from Equation (2.1), and specification of the model is completed by specifying a prior for the regression coefficients β and the disturbance variance σ^2. Definition 1 and Proposition 1 provide a recap of the familiar results on the conjugate priors and posteriors for linear regressions.

Definition 1. *The d-dimensional regression coefficients β and the disturbance variance σ^2 follow the $NIG\left(\mu, \Lambda, a, b\right)$ distribution if*

$$p\left(\beta, \sigma^2\right) \propto \left(\sigma^2\right)^{-\left(a+\frac{d}{2}+1\right)} e^{-\sigma^{-2}\left[b+\frac{1}{2}(\beta-\mu)'\Lambda(\beta-\mu)\right]}.$$

The NIG distribution is a concatenation of the multivariate normal and the inverse gamma distributions. By the marginal-conditional decomposition, we have $p(\beta, \sigma^2) = p(\beta \,|\sigma^2)\, p(\sigma^2)$, where the former is the multivariate normal distribution with the mean μ and the precision $\sigma^{-2}\Lambda$. The covariance matrix is $\sigma^2\Lambda^{-1}$. The latter is the inverse gamma distribution with the hyperparameters a and b.

Proposition 1. *If β and σ^2 have the prior $NIG(\mu, \Lambda, a, b)$, the posterior is $NIG\left(\bar{\mu}, \bar{\Lambda}, \bar{a}, \bar{b}\right)$:*

$$\bar{\mu} = (\Lambda + X'X)^{-1}(\Lambda\mu + X'Y),$$

$$\bar{\Lambda} = \Lambda + X'X,$$

$$\bar{a} = a + \frac{n}{2},$$

$$\bar{b} = b + \frac{1}{2}Y'Y + \frac{1}{2}\mu'\Lambda\mu - \frac{1}{2}\bar{\mu}'\bar{\Lambda}\bar{\mu}.$$

A non-informative prior takes the form $p(\beta, \sigma^2) \propto \sigma^{-2}$, which can be represented by an improper distribution $NIG\left(\mu_0, 0_{d\times d}, -\frac{d}{2}, 0\right)$, for an arbitrary $d \times 1$ vector μ_0. As a corollary of Proposition 1, the non-informative prior leads to the posterior $NIG\left(\tilde{\mu}, \tilde{\Lambda}, \tilde{a}, \tilde{b}\right)$, where

$$\tilde{\mu} = (X'X)^{-1}X'Y,$$

$$\tilde{\Lambda} = X'X,$$

$$\tilde{a} = \frac{n-d}{2},$$
$$\tilde{b} = \frac{1}{2}Y'Y - \frac{1}{2}Y'X\left(X'X\right)^{-1}X'Y.$$

Throughout the paper, upper-bar-form variables (e.g., $\bar{\mu}$) denote the posterior under an informative (proper) prior, while tilde-form variables (e.g., $\tilde{\mu}$) denote the posterior under a non-informative prior. They are determined by sufficient statistics $X'X$, $X'Y$ and $Y'Y$. The tilde-form posterior mean $\tilde{\mu}$ is identical to the classical ordinary least squares (OLS) estimator.

It is helpful to conceptualize conjugate distributions in terms of pseudo data. Proposition 2 shows the pseudo data extracted from the NIG distribution. Because data provide no additional information than sufficient statistics for posterior inference, pseudo data are unique up to sufficient statistics as well.

Proposition 2. *Given d-dimensional* $NIG\left(\mu, \Lambda, a, b\right)$, *where* Λ *is positive definite,* $a > 0$, $b > 0$ *and* $n = 2a + d$ *is a positive integer, we construct pseudo data such that*

$$X = \begin{pmatrix} \Lambda^{1/2} \\ 0_{(n-d)\times d} \end{pmatrix}, Y = \begin{pmatrix} \Lambda^{1/2}\mu \\ \sqrt{b/a}\cdot 1_{(n-d)\times 1} \end{pmatrix},$$

where $\Lambda^{1/2}$ *is the Cholesky factor of* Λ. *Under the non-informative prior and the pseudo data, the posterior distribution reproduces* $NIG\left(\mu, \Lambda, a, b\right)$.

2.2 Summation Operator

Motivated by the pseudo data interpretation of conjugate distributions, Qian (2018) defines the NIG summation operator. The sum of two NIG distributions can be thought as the tilde-form posterior distribution obtained by concatenating the data extracted from two NIG distributions. Alternatively, we may choose one of the NIGs as the prior, extract data from the other NIG, and run a Bayesian linear regression by Proposition 1. The upper-bar-form posterior distribution corresponds to the sum of NIGs.

Definition 2. *Consider d-dimensional* $NIG\left(\mu_1, \Lambda_1, a_1, b_1\right)$ *and* $NIG\left(\mu_2, \Lambda_2, a_2, b_2\right)$*. If a distribution* $NIG\left(\mu, \Lambda, a, b\right)$ *satisfies*

$$\mu = \left(\Lambda_1 + \Lambda_2\right)^{-1}\left(\Lambda_1\mu_1 + \Lambda_2\mu_2\right),$$

$$\Lambda = \Lambda_1 + \Lambda_2,$$

$$a = a_1 + a_2 + \frac{d}{2},$$

$$b = b_1 + b_2 + \frac{1}{2}\left(\mu_1 - \mu\right)'\Lambda_1\left(\mu_1 - \mu\right) + \frac{1}{2}\left(\mu_2 - \mu\right)'\Lambda_2\left(\mu_2 - \mu\right),$$

it is the sum of two NIG distributions, denoted by

$$NIG\left(\mu, \Lambda, a, b\right) = NIG\left(\mu_1, \Lambda_1, a_1, b_1\right) + NIG\left(\mu_2, \Lambda_2, a_2, b_2\right),$$

or more compactly,

$$NIG\left(\mu, \Lambda, a, b\right) = \sum_{j=1}^{2} NIG\left(\mu_j, \Lambda_j, a_j, b_j\right).$$

Definition 2 indicates that the NIG summation operator satisfies commutativity and associativity:

$$NIG\left(\mu_1,\Lambda_1,a_1,b_1\right)+NIG\left(\mu_2,\Lambda_2,a_2,b_2\right)=NIG\left(\mu_2,\Lambda_2,a_2,b_2\right)+NIG\left(\mu_1,\Lambda_1,a_1,b_1\right),$$

$$\sum_{j=1}^{3}NIG\left(\mu_j,\Lambda_j,a_j,b_j\right)=NIG\left(\mu_1,\Lambda_1,a_1,b_1\right)+\sum_{j=2}^{3}NIG\left(\mu_j,\Lambda_j,a_j,b_j\right).$$

As is shown in Proposition 3, a striking feature of summation is that the non-informative distribution $NIG\left(\mu_0, 0_{d\times d}, -\frac{d}{2}, 0\right)$ fulfills the role of zero in mathematics. When the NIG summation operator is applied to the non-informative distribution, the neutral element leaves unchanged any NIG distribution.

Proposition 3. *The NIG summation operator has the additive identity*

$$NIG\left(\mu,\Lambda,a,b\right)+NIG\left(\mu_0,0_{d\times d},-\frac{d}{2},0\right)=NIG\left(\mu,\Lambda,a,b\right).$$

The NIG summation operator separates the roles of data and prior information in the posterior distribution. Proposition 4 established a high-level link between Bayesian and frequentist estimators, as the former is represented by the upper-bar-form posterior $NIG\left(\bar{\mu},\bar{\Lambda},\bar{a},\bar{b}\right)$, and the latter is analogues to the tilde-form $NIG\left(\tilde{\mu},\tilde{\Lambda},\tilde{a},\tilde{b}\right)$, where $\tilde{\mu}$ is the OLS estimator. Proposition 4 may be informally read as an equation:

Bayesian estimator = prior + frequentist estimator.

Proposition 4. *Consider the prior and posterior specified in Propositions 1. We have*

$$NIG\left(\bar{\mu},\bar{\Lambda},\bar{a},\bar{b}\right)=NIG\left(\mu,\Lambda,a,b\right)+NIG\left(\tilde{\mu},\tilde{\Lambda},\tilde{a},\tilde{b}\right).$$

Qian (2018) shows that Proposition 4 facilitates regression variable selection, where mixture-NIG priors have adaptive shrinkage effects on the regression coefficients. For example, Bayesian lasso (see Tibshirani (1996) and Park and Casella (2008)) has the double exponential prior represented as a scale mixture of NIG distributions.

The regression data are also separable by the NIG summation operator. If data are divided into subsets, Proposition 5 shows that the tilde-form posterior is the sum of subset tilde-form posteriors. Because of commutativity and associativity, summation can be processed by divide and conquer. Summation results do not change with partition methods.

Proposition 5. *Partition data X, Y into k subsets X_j, Y_j, $j = 1, \ldots, k$. Let the subset tilde-form posterior distributions be $NIG\left(\tilde{\mu}_j,\tilde{\Lambda}_j,\tilde{a}_j,\tilde{b}_j\right)$ under the non-informative prior. The full-sample tilde-form posterior $NIG\left(\tilde{\mu},\tilde{\Lambda},\tilde{a},\tilde{b}\right)$ is reproduced by*

$$NIG\left(\tilde{\mu},\tilde{\Lambda},\tilde{a},\tilde{b}\right)=\sum_{j=1}^{k}NIG\left(\tilde{\mu}_j,\tilde{\Lambda}_j,\tilde{a}_j,\tilde{b}_j\right).$$

Propositions 4 and 5 justify online updating for flow data. For example, when new data X_{k+1}, Y_{k+1} come in, we make an update $NIG\left(\bar{\mu}, \bar{\Lambda}, \bar{a}, \bar{b}\right) + NIG\left(\tilde{\mu}_{k+1}, \tilde{\Lambda}_{k+1}, \tilde{a}_{k+1}, \tilde{b}_{k+1}\right)$,which equals the regression results obtained by concatenating all data points $NIG\left(\mu, \Lambda, a, b\right) + \sum_{j=1}^{k+1} NIG\left(\tilde{\mu}_j, \tilde{\Lambda}_j, \tilde{a}_j, \tilde{b}_j\right)$.

2.3 Subtraction Operator

The NIG distribution serves as a container of compressed data in the form of sufficient statistics. The summation operator adds data to the container, while the subtraction operator deducts data from the container. The NIG subtraction operator is defined by reverting Definition 2.

Definition 3. *Consider d-dimensional* $NIG\left(\mu, \Lambda, a, b\right)$ *and* $NIG\left(\mu_1, \Lambda_1, a_1, b_1\right)$. *If a distribution* $NIG\left(\mu_2, \Lambda_2, a_2, b_2\right)$ *satisfies*

$$\mu_2 = \left(\Lambda - \Lambda_1\right)^{-1}\left(\Lambda\mu - \Lambda_1\mu_1\right),$$

$$\Lambda_2 = \Lambda - \Lambda_1,$$

$$a_2 = a - a_1 - \frac{d}{2},$$

$$b_2 = b - b_1 - \frac{1}{2}\left(\mu_1 - \mu\right)'\Lambda_1\left(\mu_1 - \mu\right) - \frac{1}{2}\left(\mu_2 - \mu\right)'\Lambda_2\left(\mu_2 - \mu\right),$$

it is the difference of two NIG distributions, denoted by

$$NIG\left(\mu_2, \Lambda_2, a_2, b_2\right) = NIG\left(\mu, \Lambda, a, b\right) - NIG\left(\mu_1, \Lambda_1, a_1, b_1\right).$$

The NIG subtraction operator has broad applications. We provide three examples.

First, k-fold cross-validation. As in Proposition 5, data are partitioned into k subsets, one of which is reserved for testing the model and the remaining are used for parameter estimation. The process is repeated k times, with each subset used once as the validation data. As an alternative to repeated estimation, an efficient approach is to estimate the model once for the full-sample posterior $NIG\left(\bar{\mu}, \bar{\Lambda}, \bar{a}, \bar{b}\right)$. The NIG subtraction operation $NIG\left(\bar{\mu}, \bar{\Lambda}, \bar{a}, \bar{b}\right) - NIG\left(\tilde{\mu}_j, \tilde{\Lambda}_j, \tilde{a}_j, \tilde{b}_j\right)$ provides the results with the j^{th} subset (i.e., X_j, Y_j) excluded from the training data.

Second, rolling-window regression. Suppose that we have daily data and update regressions every month with the most recent data of $m+1$ months. Let $NIG\left(\tilde{\mu}_j, \tilde{\Lambda}_j, \tilde{a}_j, \tilde{b}_j\right)$ be the subset tilde-form posterior by the j^{th} month data alone. The dynamic posterior is given by

$$NIG\left(\bar{\mu}_k, \bar{\Lambda}_k, \bar{a}_k, \bar{b}_k\right) = NIG\left(\mu, \Lambda, a, b\right) + \sum_{j=k-m}^{k} NIG\left(\tilde{\mu}_j, \tilde{\Lambda}_j, \tilde{a}_j, \tilde{b}_j\right).$$

In the next month, the rolling-window regression is updated by adding new data and retiring old data:

$$\begin{aligned} NIG\left(\bar{\mu}_{k+1}, \bar{\Lambda}_{k+1}, \bar{a}_{k+1}, \bar{b}_{k+1}\right) = & NIG\left(\bar{\mu}_{k}, \bar{\Lambda}_{k}, \bar{a}_{k}, \bar{b}_{k}\right) \\ & + NIG\left(\tilde{\mu}_{k+1}, \tilde{\Lambda}_{k+1}, \tilde{a}_{k+1}, \tilde{b}_{k+1}\right) \\ & - NIG\left(\tilde{\mu}_{k-m}, \tilde{\Lambda}_{k-m}, \tilde{a}_{k-m}, \tilde{b}_{k-m}\right). \end{aligned}$$

Third, perfect collinearity. Suppose that the predictors include an intercept and month dummy variables. Subset regressions by the j^{th} month data alone suffer from perfect collinearity. The problem can be addressed by borrowing a prior for each subset regression:

$$NIG\left(\ddot{\mu}_j, \ddot{\Lambda}_j, \ddot{a}_j, \ddot{b}_j\right) = NIG\left(\mu, \Lambda, a, b\right) + NIG\left(\tilde{\mu}_j, \tilde{\Lambda}_j, \tilde{a}_j, \tilde{b}_j\right).$$

The borrowed priors must be repaid eventually. The m copies of duplicated priors are subtracted from the posterior:

$$NIG\left(\bar{\mu}_k, \bar{\Lambda}_k, \bar{a}_k, \bar{b}_k\right) = \sum_{j=k-m}^{k} NIG\left(\ddot{\mu}_j, \ddot{\Lambda}_j, \ddot{a}_j, \ddot{b}_j\right) - \sum_{j=1}^{m} NIG\left(\mu, \Lambda, a, b\right).$$

2.4 Multiplication Operator

In the above example, the borrowed priors amount to $\sum_{j=1}^{m} NIG\left(\mu, \Lambda, a, b\right)$, which seem to be m times $NIG\left(\mu, \Lambda, a, b\right)$. This is a motivation of the scalar multiplication for simplifying the repeated summation. Moreover, the operator is not limited to integer multiplication. If we multiply the

NIG distribution by a real number between zero and one, we interpret it as a discount factor of the NIG distribution.

Definition 4. *Consider d-dimensional* $NIG(\mu, \Lambda, a, b)$*. Let* δ *be a nonnegative scalar. If a distribution* $NIG\left(\hat{\mu}, \hat{\Lambda}, \hat{a}, \hat{b}\right)$ *satisfies*

$$\hat{\mu} = \mu,$$

$$\hat{\Lambda} = \delta\Lambda,$$

$$\hat{a} = \delta\left(a + \frac{d}{2}\right) - \frac{d}{2},$$

$$\hat{b} = \delta b,$$

it is scalar multiplication of the NIG distribution, denoted by

$$NIG\left(\hat{\mu}, \hat{\Lambda}, \hat{a}, \hat{b}\right) = \delta \cdot NIG(\mu, \Lambda, a, b).$$

Proposition 6 shows that the NIG multiplication operator follows the law of distributivity and associativity. The NIG summation operation has a zero, which is also the result of any NIG distribution multiplied by zero.

Proposition 6. *The NIG scalar multiplication operator satisfies*

1. *Distributivity:*

$$\sum_{j=1}^{k} \delta_j \cdot NIG(\mu, \Lambda, a, b) = \left(\sum_{j=1}^{k} \delta_j\right) \cdot NIG(\mu, \Lambda, a, b),$$

$$\sum_{j=1}^{k} \delta \cdot NIG\left(\mu_j, \Lambda_j, a_j, b_j\right) = \delta \cdot \sum_{j=1}^{k} NIG\left(\mu_j, \Lambda_j, a_j, b_j\right).$$

2. *Associativity:*

$$\delta_1\delta_2 \cdot NIG\left(\mu, \Lambda, a, b\right) = \delta_1 \cdot \left[\delta_2 \cdot NIG\left(\mu, \Lambda, a, b\right)\right].$$

3. *Identity/zero element:*

$$1 \cdot NIG\left(\mu, \Lambda, a, b\right) = NIG\left(\mu, \Lambda, a, b\right),$$

$$0 \cdot NIG\left(\mu, \Lambda, a, b\right) = NIG\left(\mu, 0_{d\times d}, -\frac{d}{2}, 0\right).$$

A corollary of Proposition 6 is the equivalence between scalar multiplication and repeated summation. For an integer δ, we have

$$\delta \cdot NIG\left(\mu, \Lambda, a, b\right) = \sum_{j=1}^{\delta} NIG\left(\mu, \Lambda, a, b\right),$$

because we can write $\delta = \sum_{i=1}^{\delta} 1$ and apply the distributive law of scalar multiplication.

The NIG scalar multiplication facilitates weighted regressions, in which a scalar $\delta \in [0, 1]$ serves as the discount (i.e., forgetting or smoothing) factor. For time series data, it is desirable to down-weight past observations as the time elapses and place more weights on recent observations. Just as the geometrically weighted least squares is an extension to OLS, the weighted NIG summation is an extension to the equal-weight summation.

By the NIG summation and multiplication rules, weighted summation satisfies the identity

$$\sum_{j=1}^{k} \delta^{k-j} \cdot NIG\left(\tilde{\mu}_j, \tilde{\Lambda}_j, \tilde{a}_j, \tilde{b}_j\right) = NIG\left(\hat{\mu}_k, \hat{\Lambda}_k, \hat{a}_k, \hat{b}_k\right),$$

where

$$\hat{\mu}_k = \left(\sum_{j=1}^{k} \delta^{k-j} \tilde{\Lambda}_j\right)^{-1} \left(\sum_{j=1}^{k} \delta^{k-j} \tilde{\Lambda}_j \tilde{\mu}_j\right),$$

$$\hat{\Lambda}_k = \sum_{j=1}^{k} \delta^{k-j} \tilde{\Lambda}_j,$$

$$\hat{a}_k = -\frac{d}{2} + \sum_{j=1}^{k} \delta^{k-j} \left(\tilde{a}_j + \frac{d}{2}\right),$$

$$\hat{b}_k = \sum_{j=1}^{k} \delta^{k-j} \tilde{b}_j + \frac{1}{2} \sum_{j=1}^{k} \delta^{k-j} \left(\tilde{\mu}_j - \hat{\mu}_k\right)' \tilde{\Lambda}_j \left(\tilde{\mu}_j - \hat{\mu}_k\right).$$

Because $NIG\left(\tilde{\mu}_j, \tilde{\Lambda}_j, \tilde{a}_j, \tilde{b}_j\right)$ is the posterior distribution obtained by regressing Y_j on X_j under the non-informative prior, we have $\tilde{\Lambda}_j = X_j' X_j$ and $\tilde{\Lambda}_j \tilde{\mu}_j = X_j' Y_j$. It follows that

$$\hat{\mu}_k = \left(\sum_{j=1}^{k} \delta^{k-j} X_j' X_j\right)^{-1} \left(\sum_{j=1}^{k} \delta^{k-j} X_j' Y_j\right),$$

which is recognized as the geometrically weighted least squares estimator if $\delta < 1$, and the OLS estimator if $\delta = 1$.

Weighted regressions support online update by a recursive formula:

$$NIG\left(\hat{\mu}_k, \hat{\Lambda}_k, \hat{a}_k, \hat{b}_k\right) = \delta \cdot NIG\left(\hat{\mu}_{k-1}, \hat{\Lambda}_{k-1}, \hat{a}_{k-1}, \hat{b}_{k-1}\right) + NIG\left(\tilde{\mu}_k, \tilde{\Lambda}_k, \tilde{a}_k, \tilde{b}_k\right),$$

for $k > 1$ and the starting value is $NIG\left(\hat{\mu}_1, \hat{\Lambda}_1, \hat{a}_1, \hat{b}_1\right) = NIG\left(\tilde{\mu}_1, \tilde{\Lambda}_1, \tilde{a}_1, \tilde{b}_1\right)$.

In practice, we may encounter perfect collinearity or inadequate sample size in the subset regressions. A solution is to borrow a prior $NIG\left(\mu, \Lambda, a, b\right)$ and apply the distributive law of multiplication:

$$\begin{aligned} NIG\left(\bar{\mu}_k, \bar{\Lambda}_k, \bar{a}_k, \bar{b}_k\right) &= \sum_{j=1}^{k} \delta^{k-j}\left[NIG\left(\tilde{\mu}_j, \tilde{\Lambda}_j, \tilde{a}_j, \tilde{b}_j\right) + NIG\left(\mu, \Lambda, a, b\right)\right] \\ &= NIG\left(\hat{\mu}_k, \hat{\Lambda}_k, \hat{a}_k, \hat{b}_k\right) + \frac{1-\delta^k}{1-\delta} NIG\left(\mu, \Lambda, a, b\right). \end{aligned}$$

To offset the effects of the borrowed prior, we apply the NIG subtraction:

$$NIG\left(\hat{\mu}_k, \hat{\Lambda}_k, \hat{a}_k, \hat{b}_k\right) = NIG\left(\bar{\mu}_k, \bar{\Lambda}_k, \bar{a}_k, \bar{b}_k\right) - \frac{1-\delta^k}{1-\delta} NIG\left(\mu, \Lambda, a, b\right).$$

2.5 Marginal Likelihood

An attraction of the conjugate prior is the analytically tractable marginal likelihood, which is the key to Bayesian model selection and averaging. In the context of Bayesian linear regressions, the marginal likelihood can be derived by Equation (2.1) in two steps. First, conditional on σ^2, the data Y follows the multivariate normal distribution with the mean $X\mu$ and the covariance matrix $\sigma^2 V$, where $V = I_n + X\Lambda^{-1}X'$, because the prior of β is normal with the mean μ and the covariance matrix $\sigma^2\Lambda^{-1}$. Second, marginalization over the inverse gamma distributed σ^2 provides a scale mixture that follows the multivariate t distribution, namely $t\left(X\mu, \frac{b}{a}V, 2a\right)$.

The marginal likelihood is the density of the t distribution:

$$p(Y) = \frac{\Gamma\left(a+\frac{n}{2}\right) b^a}{\Gamma(a)(2\pi)^{\frac{n}{2}}} |V|^{-\frac{1}{2}} \left[b + \frac{1}{2}(Y - X\mu)' V^{-1} (Y - X\mu)\right]^{-\left(a+\frac{n}{2}\right)}.$$

Unfortunately, the formula is unusable when the sample size is large, as it is difficult to save/load a $n \times n$ matrix in the computer memory, and it costs $O(n^3)$ operations to invert such a large matrix and compute its determinant. The NIG summation operator offers a computationally efficient method for evaluating the marginal likelihood by the Bayes formula

$$p(Y) = \frac{p(\beta, \sigma^2)\, p(Y|\beta, \sigma^2)}{p(\beta, \sigma^2 |Y)},$$

where $p(\beta, \sigma^2)$ and $p(\beta, \sigma^2|Y)$ are densities of $NIG(\mu, \Lambda, a, b)$ and $NIG\left(\bar{\mu}, \bar{\Lambda}, \bar{a}, \bar{b}\right)$ respectively. The latter can be computed by the NIG summation (Propositions 4 and 5). It is feasible to evaluate the densities by divide and conquer even if the sample size is large. Note that the NIG densities can be evaluated at an arbitrary β and σ^2 without changing the marginal likelihood results. For example, we may evaluate the densities at $\beta = 0$, $\sigma^2 = 1$ to simplify the functional forms:

$$p\left(\beta, \sigma^2\right) = (2\pi)^{-\frac{d}{2}} |\Lambda|^{\frac{1}{2}} e^{-\frac{1}{2}\mu'\Lambda\mu} \frac{b^a}{\Gamma(a)} e^{-b},$$

$$p\left(\beta, \sigma^2 |Y\right) = (2\pi)^{-\frac{d}{2}} \left|\bar{\Lambda}\right|^{\frac{1}{2}} e^{-\frac{1}{2}\bar{\mu}'\bar{\Lambda}\bar{\mu}} \frac{\bar{b}^{\bar{a}}}{\Gamma(\bar{a})} e^{-\bar{b}},$$

$$p\left(Y \middle| \beta, \sigma^2\right) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}Y'Y}.$$

The above expressions are fully determined by the sufficient statistics $X'X$, $X'Y$ and $Y'Y$, so the computational complexity is $O(d^2n)$, assuming that $n > d$.

2.6 Prediction

Assume that data in the forecast horizon are generated like Equation (2.1):

$$Y_f = X_f\beta + \sigma\varepsilon_f.$$

Bayesian prediction addresses the posterior predictive distribution of the response variables Y_f, conditional on the observed data Y. The posterior predictive distribution is the weighted average of the predicted likelihood function, weighted by the posterior density of parameters:

$$p(Y_f | Y) = \int\int p\left(\beta, \sigma^2 | Y\right) p\left(Y_f \left| \beta, \sigma^2, Y\right.\right) d\beta d\sigma^2.$$

Since the unknown parameters are integrated out, Bayesian prediction incorporates parameter uncertainty.

The posterior predictive distribution is analytically tractable if we specify the NIG conjugate prior. By Proposition 1, $p(\beta, \sigma^2 | Y)$ follows $NIG\left(\bar{\mu}, \bar{\Lambda}, \bar{a}, \bar{b}\right)$. Similar to derivation of the marginal likelihood, the posterior predictive distribution is the multivariate $t\left(X_f\bar{\mu}, \frac{\bar{b}}{\bar{a}}V_f, 2\bar{a}\right)$:

$$p\left(Y_f\left|Y\right.\right) = \frac{\Gamma\left(\frac{2\bar{a}+n_f}{2}\right)\bar{b}^{\bar{a}}}{\Gamma\left(\bar{a}\right)\left(2\pi\right)^{\frac{n_f}{2}}}\left|V_f\right|^{-\frac{1}{2}}\left[\bar{b}+\frac{1}{2}\left(Y_f - X_f\bar{\mu}\right)' V_f^{-1}\left(Y_f - X_f\bar{\mu}\right)\right]^{-\frac{2\bar{a}+n_f}{2}},$$

where $V_f = I + X_f\bar{\Lambda}^{-1}X_f'$ and n_f is the length of Y_f.

The NIG summation operator offers an alternative method of evaluating the posterior predictive density. The Bayes formula indicates that

$$p\left(Y_f\left|Y\right.\right) = \frac{p\left(\beta, \sigma^2\left|Y\right.\right)p\left(Y_f\left|\beta, \sigma^2, Y\right.\right)}{p\left(\beta, \sigma^2\left|Y_f, Y\right.\right)},$$

where $p\left(\beta, \sigma^2\left|Y\right.\right)$ is the density of $NIG\left(\bar{\mu}, \bar{\Lambda}, \bar{a}, \bar{b}\right)$, while $p\left(\beta, \sigma^2\left|Y_f, Y\right.\right)$ corresponds to the posterior distribution by concatenating Y and Y_f. By Propositions 4 and 5, it can be computed by the NIG summation $NIG\left(\bar{\mu}, \bar{\Lambda}, \bar{a}, \bar{b}\right)+ NIG\left(\tilde{\mu}_f, \tilde{\Lambda}_f, \tilde{a}_f, \tilde{b}_f\right)$,where $NIG\left(\tilde{\mu}_f, \tilde{\Lambda}_f, \tilde{a}_f, \tilde{b}_f\right)$ is the subset posterior obtained by using X_f, Y_f under the non-informative prior.

3. Exponential Family

The NIG summation operator can be extended to conjugate models in the exponential family. This section provides some examples.

The beta-binomial model is a common entry point for Bayesian statistics textbooks. See Gelman et al. (2014, p.29), Bolstad and Curran (2017, p.149), among others. To estimate the success probability θ, we use binomial data with y successes out of n Bernoulli trials. The likelihood is

$$p\left(y\,|\theta\right) \propto \theta^{y}\left(1-\theta\right)^{n-y}.$$

Under the conjugate prior $Beta\left(\alpha,\beta\right)$ such that

$$p\left(\theta\right) \propto \theta^{\alpha-1}\left(1-\theta\right)^{\beta-1},$$

the posterior is $Beta\left(\bar{\alpha},\bar{\beta}\right)$, where $\bar{\alpha}=\alpha+y$ and $\bar{\beta}=\beta+n-y$.

Gelman et al. (2014, p.35) give a pseudo-data interpretation of the prior $Beta\left(\alpha,\beta\right)$: "Comparing $p\left(\theta\right)$ and $p\left(y\,|\theta\right)$ suggests that this prior density is equivalent to $\alpha-1$ prior successes and $\beta-1$ prior failures."

Definition 5 provides the summation rules for beta distributions. The sum of beta distributions can be viewed as the posterior distribution obtained by concatenating the pseudo data extracted from two beta distributions under the uniform prior.

Definition 5. *Consider beta distributions* $Beta\left(\alpha_1,\beta_1\right)$ *and* $Beta\left(\alpha_2,\beta_2\right)$*. If a distribution* $Beta\left(\alpha,\beta\right)$ *satisfies*

$$\alpha=\alpha_1+\alpha_2-1,$$

$$\beta=\beta_1+\beta_2-1,$$

it is the sum of beta distributions, denoted by

$$Beta\left(\alpha,\beta\right)=Beta\left(\alpha_1,\beta_1\right)+Beta\left(\alpha_2,\beta_2\right).$$

The beta summation operator satisfies commutativity and associativity. Most importantly, we have the additive identity:

$$Beta\,(\alpha, \beta) = Beta\,(1, 1) + Beta\,(\alpha, \beta)\,,$$

where $Beta\,(1, 1)$ is the uniform distribution that plays the role of zero in summation. The uniform prior contains zero pseudo data (i.e., 0 success and 0 failure). Adding $Beta\,(1, 1)$ to $Beta\,(\alpha, \beta)$ leaves the latter unchanged.

The posterior is a compromise between data and prior information, as the posterior mean is a weighted average of the prior mean and the sample proportion (Gelman et al., 2014, p.32). The beta summation operator connects the prior, data and posterior by a simple equation:

$$Beta\left(\bar{\alpha}, \bar{\beta}\right) = Beta\,(\alpha, \beta) + Beta\left(\tilde{\alpha}, \tilde{\beta}\right),$$

where $Beta\left(\bar{\alpha}, \bar{\beta}\right)$ is the posterior under the prior $Beta\,(\alpha, \beta)$, and $Beta\left(\tilde{\alpha}, \tilde{\beta}\right)$ is the posterior under the uniform prior.

The gamma-Poisson model is another example of conjugacy in Bayesian statistics textbooks. See Gelman et al. (2014, p.43), Bolstad and Curran (2017, p.193). Suppose that $Y_i, i = 1, \ldots, n$ follow the Poisson distribution with the intensity parameter θ. The likelihood is

$$p\,(Y\,|\theta\,) \propto \theta^{\sum_{i=1}^{n} Y_i} e^{-n\theta}.$$

Under the conjugate prior $Gamma\,(A, B)$ such that

$$p\,(\theta) \propto \theta^{A-1} e^{-B\theta},$$

the posterior is $Gamma\left(\bar{A}, \bar{B}\right)$, where $\bar{A} = A + \sum_{i=1}^{n} Y_i$ and $\bar{B} = B + n$.

The summation rules for gamma distributions are similar to those in NIG and beta distributions, as is shown in Definition 6.

Definition 6. *Consider gamma distributions* $Gamma\,(A_1, B_1)$ *and* $Gamma\,(A_2, B_2)$*. If a distribution* $Gamma\,(A, B)$ *satisfies*

$$A = A_1 + A_2 - 1,$$
$$B = B_1 + B_2,$$

it is the sum of gamma distributions, denoted by

$$Gamma\,(A, B) = Gamma\,(A_1, B_1) + Gamma\,(A_2, B_2)\,.$$

Commutativity and associativity still hold for the gamma summation operator. In particular, we have the additive identity:

$$Gamma\,(A, B) = Gamma\,(1, 0) + Gamma\,(A, B)\,,$$

where $Gamma\,(1, 0)$ is the positive uniform prior $p\,(\theta) \propto 1$, $\theta > 0$.

Diaconis and Ylvisaker (1979) analyze conjugate priors for the exponential family. Let $Y_i, i = 1, \ldots, n$ be random samples from an exponential family distribution such that

$$p\left(Y_i\left|\theta\right.\right)=f\left(Y_i\right)g\left(\theta\right)e^{\phi(\theta)'T(Y_i)}.$$

Given the sufficient statistics $T\left(\cdot\right)$, the natural parameterization $\phi\left(\cdot\right)$ and the normalization factor $g\left(\cdot\right)$, Gelman et al. (2014, p.37) and Bernardo and Smith (2000, p.266) show that the conjugate prior distribution, denoted by $\epsilon_{T\phi g}\left(\tau,\upsilon\right)$, takes the form

$$p\left(\theta\right)\propto\left[g\left(\theta\right)\right]^{\upsilon}e^{\phi(\theta)'\tau}.$$

The posterior distribution $\epsilon_{T\phi g}\left(\bar{\tau},\bar{\upsilon}\right)$ is obtained by adding the hyperparameter τ to the sufficient statistics and adding the hyperparameter υ to the sample size:

$$\bar{\tau}=\tau+\sum_{i=1}^{n}T\left(Y_i\right),$$

$$\bar{\upsilon}=\upsilon+n.$$

If we specify $\tau=0$ and $\upsilon=0$, we have a non-informative prior $p\left(\theta\right)\propto 1$. Definition 7 provides the summation rules for the exponential family, and Proposition 7 is an extension of Proposition 4.

Definition 7. *Let* $\epsilon_{T\phi g}\left(\tau_0,\upsilon_0\right)$ *be a non-informative prior. Consider distri-*

butions $\epsilon_{T\phi g}(\tau_1, v_1)$ *and* $\epsilon_{T\phi g}(\tau_2, v_2)$. *If a distribution* $\epsilon_{T\phi g}(\tau, v)$ *satisfies*

$$\tau = \tau_1 + \tau_2 - \tau_0,$$

$$v = v_1 + v_2 - v_0,$$

it is the sum of two distributions, denoted by

$$\epsilon_{T\phi g}(\tau, v) = \epsilon_{T\phi g}(\tau_1, v_1) + \epsilon_{T\phi g}(\tau_2, v_2).$$

Proposition 7. *Let* $\epsilon_{T\phi g}(\bar{\tau}, \bar{v})$ *be the posterior under a conjugate prior* $\epsilon_{T\phi g}(\tau, v)$. *Let* $\epsilon_{T\phi g}(\tilde{\tau}, \tilde{v})$ *be the posterior under a non-informative prior* $\epsilon_{T\phi g}(\tau_0, v_0)$ *with the same data. We have*

$$\epsilon_{T\phi g}(\bar{\tau}, \bar{v}) = \epsilon_{T\phi g}(\tau, v) + \epsilon_{T\phi g}(\tilde{\tau}, \tilde{v}).$$

Definition 7 is compatible with Definition 6. For the gamma-Poisson model, we have $T(Y_i) = Y_i$, $\phi(\theta) = \ln\theta$ and $g(\theta) = e^{-\theta}$. The conjugate prior takes the form $p(\theta) \propto \theta^{\tau} e^{-v\theta}$, which is $Gamma(\tau + 1, v)$. We specify $\tau_0 = 0$ and $v_0 = 0$ for a non-informative prior. Definitions 6 and 7 describe the same summation rules under different parametrization of the gamma distribution: $A = \tau + 1$ and $B = v$.

Similarly, Definition 7 is also compatible with Definition 5. We rewrite the binomial data as n Bernoulli samples: $y = \sum_{i=1}^{n} Y_i$. We have $T(Y_i) = Y_i$, $\phi(\theta) = \ln\frac{\theta}{1-\theta}$ and $g(\theta) = 1 - \theta$. The conjugate prior takes the form

$p(\theta) \propto \theta^{\tau}(1-\theta)^{v-\tau}$, which is $Beta(\tau+1, v-\tau+1)$. We specify $\tau_0 = 0$ and $v_0 = 0$ for a non-informative prior. Definitions 5 and 7 describe the same summation rules under different parametrization of the beta distribution: $\alpha = \tau + 1$ and $\beta = v - \tau + 1$.

4. An Application of Rolling-window Regressions

As the economic environment changes over time, model parameters can be time varying. Rolling-window regressions are popular in financial time series analysis. We illustrate the NIG summation and subtraction operations in rolling-window regressions based on the five-factor asset pricing model of Fama and French (2015). The factors capture the effects of size, value, profitability, investment, and market excess returns. The response variables are daily returns of five industry portfolios from July 1963 to June 2023. Data are publicly available at the website of Professor French.

We specify 10-year rolling windows with 2520 observations for each regression. The first window ranges from July 1963 to June 1973, and the second window moves forward by one month (from August 1963 to July 1973), and so on. There are 600 rolling-window regressions in total. Under the non-informative prior, the posterior mean is the OLS estimator. Figure 1 plots the estimated "beta" coefficients for the industry portfolios.

A series of regressions can be efficiently implemented by adding new data with the NIG summation and retiring old data with the NIG subtraction. As the estimation window moves forward by one month (21 observations), the added computing cost is about 6000 floating-point operations (flops). In contract, each OLS estimation with 2520 observations needs about 210000 flops, which is over 30 times higher than the computing cost by the NIG arithmetic.

5. Conclusion

A common critique of Bayesian statistics is the subjective prior in statistical inference, as the prior and posterior reflect subjective states of knowledge (Gelman, 2008). The NIG summation operator connects the subjective and objective estimators by a simple equation shown in Proposition 4. To disentangle the prior information from the posterior, the subtraction operator can be used: $NIG\left(\bar{\mu}, \bar{\Lambda}, \bar{a}, \bar{b}\right) - NIG\left(\mu, \Lambda, a, b\right)$. To reduce the prior strength by a half, the multiplication operator provides a convenient solution: $NIG\left(\bar{\mu}, \bar{\Lambda}, \bar{a}, \bar{b}\right) - \frac{1}{2} \cdot NIG\left(\mu, \Lambda, a, b\right)$.

From a pedagogical perspective, the paper shows that the hard work of manipulating density functions of conjugate distributions can be simplified by intuitive math operators: summation, subtraction, and multiplication.

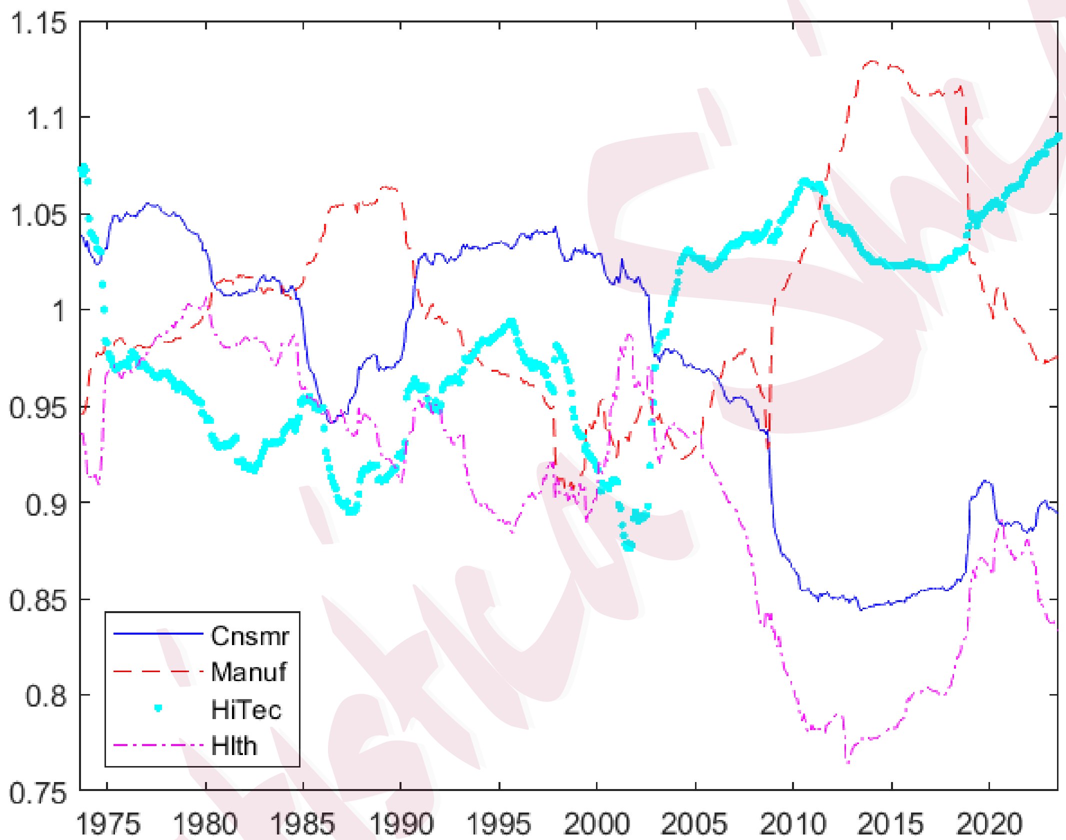

Figure 1: Rolling-window regressions of Fama-French five-factor asset pricing model. Estimated "beta" coefficients are plotted for industry portfolios.

Appendix

Proof of Proposition 1

The posterior density takes the form

$$p\left(\beta,\sigma^2\left|Y\right.\right)\propto\left(\sigma^2\right)^{-\left(a+\frac{n+d}{2}+1\right)}e^{-\sigma^{-2}\left[b+\frac{1}{2}(\beta-\mu)'\Lambda(\beta-\mu)+\frac{1}{2}(Y-X\beta)'(Y-X\beta)\right]}.$$

By completing the squares, we have

$$p\left(\beta,\sigma^2\left|Y\right.\right)\propto\left(\sigma^2\right)^{-\left(a+\frac{n+d}{2}+1\right)}e^{-\frac{1}{2}\sigma^{-2}\left[2b+(\beta-\bar{\mu})'\bar{\Lambda}(\beta-\bar{\mu})-\bar{\mu}'\bar{\Lambda}\bar{\mu}+\mu'\Lambda\mu+Y'Y\right]},$$

where $\bar{\mu}=\left(\Lambda+X'X\right)^{-1}\left(\Lambda\mu+X'Y\right)$ and $\bar{\Lambda}=\Lambda+X'X$. We recognize that it is the density of $NIG\left(\bar{\mu},\bar{\Lambda},\bar{a},\bar{b}\right)$.

Proof of Proposition 2

The pseudo data provide the following sufficient statistics: $X'X=\left(\Lambda^{1/2}\right)'\Lambda^{1/2}=\Lambda$, $X'Y=\Lambda\mu$ and $Y'Y=\mu'\Lambda\mu+2b$. Under the non-informative prior and the pseudo data, the posterior distribution is $NIG\left(\tilde{\mu},\tilde{\Lambda},\tilde{a},\tilde{b}\right)$, where

$$\tilde{\mu}=\left(X'X\right)^{-1}X'Y=\mu,$$

$$\tilde{\Lambda}=X'X=\Lambda,$$

$$\tilde{a}=\frac{n-d}{2}=a,$$

$$\tilde{b}=\frac{1}{2}Y'Y-\frac{1}{2}Y'X\left(X'X\right)^{-1}X'Y=b,$$

that is, $NIG\left(\tilde{\mu},\tilde{\Lambda},\tilde{a},\tilde{b}\right)=NIG\left(\mu,\Lambda,a,b\right)$.

Proof of Proposition 3

The additive identity is immediate from the definition of NIG summation.

Proof of Proposition 4

Under the non-informative prior, we have $\tilde{\mu} = (X'X)^{-1} X'Y$, $\tilde{\Lambda} = X'X$, $\tilde{a} = \frac{n-d}{2}$, $\tilde{b} = \frac{1}{2}Y'Y - \frac{1}{2}\tilde{\mu}'\tilde{\Lambda}\tilde{\mu}$. Denote $NIG\left(\hat{\mu}, \hat{\Lambda}, \hat{a}, \hat{b}\right) \equiv NIG\left(\mu, \Lambda, a, b\right) + NIG\left(\tilde{\mu}, \tilde{\Lambda}, \tilde{a}, \tilde{b}\right)$. By the NIG summation rules, we have

$$\hat{\mu} = (\Lambda + X'X)^{-1} (\Lambda\mu + X'Y),$$

$$\hat{\Lambda} = \Lambda + X'X,$$

$$\hat{a} = a + \frac{n}{2},$$

$$\hat{b} = b + \frac{1}{2}Y'Y + \frac{1}{2}\mu'\Lambda\mu - \frac{1}{2}\hat{\mu}'\hat{\Lambda}\hat{\mu}.$$

It follows that $NIG\left(\hat{\mu}, \hat{\Lambda}, \hat{a}, \hat{b}\right) = NIG\left(\bar{\mu}, \bar{\Lambda}, \bar{a}, \bar{b}\right)$.

Proof of Proposition 5

The subset tilde-form posterior satisfies $\tilde{\mu}_j = \left(X_j'X_j\right)^{-1} X_j'Y_j$, $\tilde{\Lambda}_j = X_j'X_j$, $\tilde{a}_j = \frac{n_j-d}{2}$, $\tilde{b}_j = \frac{1}{2}Y_j'Y_j - \frac{1}{2}\tilde{\mu}_j'\tilde{\Lambda}_j\tilde{\mu}_j$.

Note that $X'X = \sum_{j=1}^k X_j'X_j$, $X'Y = \sum_{j=1}^k X_j'Y_j$, $Y'Y = \sum_{j=1}^k Y_j'Y_j$, $n = \sum_{j=1}^k n_j$. By the NIG summation rules, we have $NIG\left(\tilde{\mu}, \tilde{\Lambda}, \tilde{a}, \tilde{b}\right) = \sum_{j=1}^k NIG\left(\tilde{\mu}_j, \tilde{\Lambda}_j, \tilde{a}_j, \tilde{b}_j\right)$.

Proof of Proposition 6

We will show the case of $n = 2$. Results for $n > 2$ can be obtained by induction. To show distributivity,

$$\delta_1 \cdot NIG(\mu, \Lambda, a, b) + \delta_2 \cdot NIG(\mu, \Lambda, a, b)$$

$$= NIG\left[\mu, \delta_1\Lambda, \delta_1\left(a + \tfrac{d}{2}\right) - \tfrac{d}{2}, \delta_1 b\right] + NIG\left[\mu, \delta_2\Lambda, \delta_2\left(a + \tfrac{d}{2}\right) - \tfrac{d}{2}, \delta_2 b\right]$$

$$= NIG\left[\mu, (\delta_1 + \delta_2)\Lambda, (\delta_1 + \delta_2)\left(a + \tfrac{d}{2}\right) - \tfrac{d}{2}, (\delta_1 + \delta_2) b\right]$$

$$= (\delta_1 + \delta_2) \cdot NIG(\mu, \Lambda, a, b)$$

$$\delta \cdot NIG(\mu_1, \Lambda_1, a_1, b_1) + \delta \cdot NIG(\mu_2, \Lambda_2, a_2, b_2)$$

$$= NIG\left[\mu_1, \delta\Lambda_1, \delta\left(a_1 + \tfrac{d}{2}\right) - \tfrac{d}{2}, \delta b_1\right] + NIG\left[\mu_2, \delta\Lambda_2, \delta\left(a_2 + \tfrac{d}{2}\right) - \tfrac{d}{2}, \delta b_2\right]$$

$$= NIG\left[\mu, \delta\Lambda, \delta\left(a + \tfrac{d}{2}\right) - \tfrac{d}{2}, \delta b\right]$$

$$= \delta \cdot NIG(\mu, \Lambda, a, b).$$

We have $\mu = (\Lambda_1 + \Lambda_2)^{-1}(\Lambda_1\mu_1 + \Lambda_2\mu_2)$, $\Lambda = \Lambda_1 + \Lambda_2$, $a = a_1 + a_2 + \frac{d}{2}$, $b = b_1 + b_2 + \frac{1}{2}(\mu_1 - \mu)'\Lambda_1(\mu_1 - \mu) + \frac{1}{2}(\mu_2 - \mu)'\Lambda_2(\mu_2 - \mu)$. By Definition 2, this is the sum of two NIGs.

To show associativity,

$$\delta_1 \cdot [\delta_2 \cdot NIG(\mu, \Lambda, a, b)]$$

$$= \delta_1 \cdot \left[NIG\left(\mu, \delta_2\Lambda, \delta_2\left(a + \tfrac{d}{2}\right) - \tfrac{d}{2}, \delta_2 b\right)\right]$$

$$= NIG\left(\mu, \delta_1\delta_2\Lambda, \delta_1\delta_2\left(a + \tfrac{d}{2}\right) - \tfrac{d}{2}, \delta_1\delta_2 b\right)$$

$$= \delta_1\delta_2 \cdot NIG(\mu, \Lambda, a, b).$$

To show identity/zero element,

$$1\cdot NIG\left(\mu,\Lambda,a,b\right)=NIG\left(\mu,1\cdot\Lambda,1\cdot\left(a+\tfrac{d}{2}\right)-\tfrac{d}{2},1\cdot b\right)=NIG\left(\mu,\Lambda,a,b\right).$$

$$0\cdot NIG\left(\mu,\Lambda,a,b\right)=NIG\left(\mu,0\cdot\Lambda,0\cdot\left(a+\tfrac{d}{2}\right)-\tfrac{d}{2},0\cdot b\right)=NIG\left(\mu,0_{d\times d},-\tfrac{d}{2},0\right).$$

Proof of Proposition 7

The posterior under the conjugate prior is obtained by $\bar{\tau}=\tau+\sum_{i=1}^{n}T\left(Y_i\right)$ and $\bar{v}=v+n$. Similarly, the posterior under the non-informative prior is $\tilde{\tau}=\tau_0+\sum_{i=1}^{n}T\left(Y_i\right)$ and $\tilde{v}=v_0+n$. By the summation rules,

$$\epsilon_{T\phi g}\left(\tau,v\right)+\epsilon_{T\phi g}\left(\tilde{\tau},\tilde{v}\right)=\epsilon_{T\phi g}\left(\tau+\tilde{\tau}-\tau_0,v+\tilde{v}-v_0\right)=\epsilon_{T\phi g}\left(\bar{\tau},\bar{v}\right).$$

References

Bernardo, J. and A. Smith (2000). *Bayesian Theory*. Chichester: John Wiley and Sons.

Bolstad, W. and J. Curran (2017). *Introduction to Bayesian Statistics (Third Edition)*. Hoboken: Wiley.

Christensen, R., W. Johnson, A. Branscum, and T. Hanson (2011). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. Boca Raton: CRC Press.

Diaconis, P. and D. Ylvisaker (1979). Conjugate priors for exponential families. *Annals of Statistics 7*, 269–281.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*, 1–22.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis 3*, 445–449.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2014). *Bayesian Data Analysis (Third Edition)*. Boca Raton: CRC Press.

Koop, G. (2003). *Bayesian Econometrics*. Hoboken: Wiley.

Park, T. and G. Casella (2008). The Bayesian Lasso. *Journal of the American Statistical Association 103*(482), 681–686.

Qian, H. (2018). Big data Bayesian linear regression and variable selection by normal-inverse-gamma summation. *Bayesian Analysis 13*(4), 1011–1035.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society B 58*, 267–288.