# The Matrix-$F$ Prior for Estimating and Testing Covariance Matrices

Joris Mulder & Luis R. Pericchi

Department of Methodology & Statistics
Tilburg University, the Netherlands

CWI talk 2018, Amsterdam, 5-4-18

# Outline

1. Problems with inverse gamma priors
2. Introducing the univariate $F$ and matrix-$F$ prior
3. The matrix-$F$ prior in regularized regression
4. The matrix-$F$ prior for testing covariance matrices
   - Testing a precise hypothesis
   - Testing inequality constrained hypotheses
5. The matrix-$F$ prior for modeling random effects covariance matrices
6. Summary

# Outline

# Modeling variance components

- The **inverse gamma prior** is the default choice for modeling variance components,

$$\sigma^2 \sim \mathcal{IG}(\alpha, \beta),$$

with prior shape parameter $\alpha$ and prior scale parameter $\beta$.

# Modeling variance components

- The **inverse gamma prior** is the default choice for modeling variance components,

$$\sigma^2 \sim \mathcal{IG}(\alpha, \beta),$$

  with prior shape parameter $\alpha$ and prior scale parameter $\beta$.
- The inverse gamma prior is conjugate for a variance of a normal population.

# Modeling variance components

- The **inverse gamma prior** is the default choice for modeling variance components,

$$\sigma^2 \sim \mathcal{IG}(\alpha, \beta),$$

with prior shape parameter $\alpha$ and prior scale parameter $\beta$.

- The inverse gamma prior is conjugate for a variance of a normal population.
- Default choice: $\alpha = \beta = \epsilon > 0$, with $\epsilon$ small, e.g., .001.

# Modeling variance components

- The **inverse gamma prior** is the default choice for modeling variance components,

$$\sigma^2 \sim \mathcal{IG}(\alpha, \beta),$$

  with prior shape parameter $\alpha$ and prior scale parameter $\beta$.

- The inverse gamma prior is conjugate for a variance of a normal population.
- Default choice: $\alpha = \beta = \epsilon > 0$, with $\epsilon$ small, e.g., .001.
- The inverse gamma prior is a proper neighboring prior of the popular Jeffreys prior $\sigma^{-2}$. Let

$$\begin{aligned} p^N(\sigma^2|\mathbf{x}) &\propto \sigma^{-2} f(\mathbf{x}|\sigma^2) \\ p(\sigma^2|\mathbf{x}) &\propto \mathcal{IG}(\sigma^2; \epsilon, \epsilon) f(\mathbf{x}|\sigma^2), \end{aligned}$$

  then

$$p(\sigma^2|\mathbf{x}) \rightarrow p^N(\sigma^2|\mathbf{x}), \text{ as } \epsilon \rightarrow 0.$$

# Problems with the inverse gamma prior

- Surprisingly, the inverse gamma can unduly be highly informative as a prior for the random effects variance in a hierarchical model,

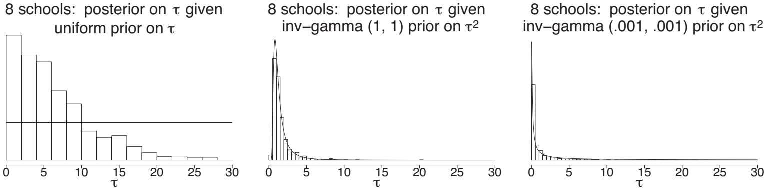
$$\begin{aligned} i\text{-th observation in group } j\text{:} \quad & y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2) \\ \text{random mean of group } j\text{:} \quad & \mu_j \sim \mathcal{N}(\mu, \tau^2). \end{aligned}$$

# Problems with the inverse gamma prior

- Surprisingly, the inverse gamma can unduly be highly informative as a prior for the random effects variance in a hierarchical model,

$$\begin{aligned} i\text{-th observation in group } j: \quad & y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2) \\ \text{random mean of group } j: \quad & \mu_j \sim \mathcal{N}(\mu, \tau^2). \end{aligned}$$

- The 8 schools example of Gelman (2006) showed the effect of the inverse gamma prior on $\tau^2$:



# Outline

# The $F$ prior

- The issue of the inverse gamma prior can be resolved by mixing the scale parameter with a gamma distribution. This results in a univariate $F$ prior:

$$\mathcal{F}(\sigma^2;\nu,\delta,b)=\int\mathcal{IG}(\sigma^2;\tfrac{\delta}{2},\psi^2)\times\mathcal{G}(\psi^2;\tfrac{\nu}{2},b^{-1})d\psi^2,$$

with degrees of freedom parameters $\nu$ and $\delta$, and scale parameter $b$.

# The $F$ prior

- The issue of the inverse gamma prior can be resolved by mixing the scale parameter with a gamma distribution. This results in a univariate $F$ prior:

$$\mathcal{F}(\sigma^2;\nu,\delta,b) = \int \mathcal{IG}(\sigma^2;\tfrac{\delta}{2},\psi^2)\times\mathcal{G}(\psi^2;\tfrac{\nu}{2},b^{-1})d\psi^2,$$

  with degrees of freedom parameters $\nu$ and $\delta$, and scale parameter $b$.
- Mixing a hyperparameter with another distribution is a way to robustify a prior.
  - **Example:** The Student $t$ prior is known to be more robust than a normal prior for regression analysis.
  - The Student $t$ prior is obtained by mixing the variance of a normal prior:

$$t(\beta;\mu,\gamma,\nu) = \int \mathcal{N}(\beta;\mu,\sigma^2)\mathcal{IG}(\sigma^2;\tfrac{\nu}{2},\tfrac{\gamma}{2})d\sigma^2.$$

# The $F$ prior

- Setting $\nu = 1$, the standard deviation has a half-$t$ distribution:

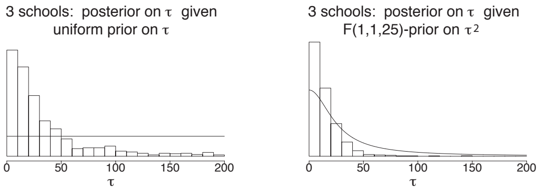
$$p(\sigma|\nu = 1, \delta, b) = \frac{2\Gamma(\frac{\delta+1}{2})}{\Gamma(\frac{\delta}{2})\sqrt{b\pi}} \left(1 + \frac{\sigma^2}{b}\right)^{\frac{\delta+1}{2}}.$$

# The $F$ prior

- Setting $\nu = 1$, the standard deviation has a half-$t$ distribution:

$$p(\sigma|\nu = 1, \delta, b) = \frac{2\Gamma(\frac{\delta+1}{2})}{\Gamma(\frac{\delta}{2})\sqrt{b\pi}} \left(1 + \frac{\sigma^2}{b}\right)^{\frac{\delta+1}{2}} .$$

- The $F$ prior results in more desirable behavior than the inverse gamma prior for school data (Gelman, 2006).



# The matrix-$F$ prior

- In a multivariate setting, the **inverse Wishart prior** is the default choice for a $k \times k$ covariance matrix.

# The matrix-$F$ prior

- In a multivariate setting, the **inverse Wishart prior** is the default choice for a $k \times k$ covariance matrix.
- The inverse Wishart prior is a matrix generalization of the inverse gamma prior, and thus has similar issues.

# The matrix-$F$ prior

- In a multivariate setting, the **inverse Wishart prior** is the default choice for a $k \times k$ covariance matrix.
- The inverse Wishart prior is a matrix generalization of the inverse gamma prior, and thus has similar issues.
- We propose to robustify the inverse Wishart by **mixing the scale matrix with a Wishart distribution**:

$$\mathcal{F}(\boldsymbol{\Sigma}; \nu, \delta, \mathbf{S}) = \int \mathcal{IW}(\boldsymbol{\Sigma}; \delta + k - 1, \boldsymbol{\Psi}) \times \mathcal{W}(\boldsymbol{\Psi}; \nu, \mathbf{B}) d\boldsymbol{\Psi},$$

where $\nu$ controls the behavior near the origin of $|\boldsymbol{\Sigma}|$, $\delta$ controls the behavior in the tails of $|\boldsymbol{\Sigma}|$, and $\mathbf{B}$ is a scale matrix.

# The matrix-$F$ prior

- In a multivariate setting, the **inverse Wishart prior** is the default choice for a $k \times k$ covariance matrix.
- The inverse Wishart prior is a matrix generalization of the inverse gamma prior, and thus has similar issues.
- We propose to robustify the inverse Wishart by **mixing the scale matrix with a Wishart distribution**:

$$\mathcal{F}(\boldsymbol{\Sigma}; \nu, \delta, \mathbf{S}) = \int \mathcal{IW}(\boldsymbol{\Sigma}; \delta + k - 1, \boldsymbol{\Psi}) \times \mathcal{W}(\boldsymbol{\Psi}; \nu, \mathbf{B}) d\boldsymbol{\Psi},$$

  where $\nu$ controls the behavior near the origin of $|\boldsymbol{\Sigma}|$, $\delta$ controls the behavior in the tails of $|\boldsymbol{\Sigma}|$, and $\mathbf{B}$ is a scale matrix.
- Setting $\mathbf{S} = \mathbf{I}_k$ yields the standard matrix-$F$ distribution (Dawid, 1981).

# Properties of the matrix-$F$ distribution

- **Reciprocity:**

$$\mathbf{\Sigma} \sim \mathcal{F}(\nu, \delta, \mathbf{S}) \Rightarrow \mathbf{\Sigma}^{-1} \sim \mathcal{F}(\delta + k - 1, \nu - k + 1, \mathbf{S}^{-1})$$

# Properties of the matrix-$F$ distribution

- **Reciprocity:**

$$\boldsymbol{\Sigma} \sim \mathcal{F}(\nu, \delta, \mathbf{S}) \Rightarrow \boldsymbol{\Sigma}^{-1} \sim \mathcal{F}(\delta + k - 1, \nu - k + 1, \mathbf{S}^{-1})$$

- **Invariant under marginalization:**

$$\boldsymbol{\Sigma} \sim \mathcal{F}(\nu, \delta, \mathbf{S}) \Rightarrow \boldsymbol{\Sigma}_{11} \sim \mathcal{F}(\nu, \delta, \mathbf{S}_{11})$$

# Properties of the matrix-$F$ distribution

- **Reciprocity:**

$$\mathbf{\Sigma} \sim \mathcal{F}(\nu, \delta, \mathbf{S}) \Rightarrow \mathbf{\Sigma}^{-1} \sim \mathcal{F}(\delta + k - 1, \nu - k + 1, \mathbf{S}^{-1})$$

- **Invariant under marginalization:**

$$\mathbf{\Sigma} \sim \mathcal{F}(\nu, \delta, \mathbf{S}) \Rightarrow \mathbf{\Sigma}_{11} \sim \mathcal{F}(\nu, \delta, \mathbf{S}_{11})$$

- **Implementation in Gibbs sampler:**
  The matrix-$F$ prior can easily be implemented in a Gibbs sampler using a parameter expansion:

$$\mathbf{\Sigma} \sim \mathcal{F}(\nu, \delta, \mathbf{S}) \Leftrightarrow \begin{cases} \mathbf{\Sigma} \sim \mathcal{IW}(\mathbf{\Sigma}; \delta + k - 1, \mathbf{\Psi}) \\ \mathbf{\Psi} \sim \mathcal{W}(\mathbf{\Psi}; \nu, \mathbf{B}) \end{cases}$$

  Then

$$\mathbf{\Psi} | \mathbf{\Sigma} \sim \mathcal{W}(\nu + \delta + k - 1, (\mathbf{B}^{-1} + \mathbf{\Sigma}^{-1})^{-1}).$$

# Properties of the matrix-$F$ distribution

- **Implementation in R:**
  - In R, draw $\mathbf{\Sigma}$ having an **inverse Wishart prior**:

    ```
    Sigma <- solve(rwish(v=n+k,S=solve(SS + B0))
    ```

  - In R, draw $\mathbf{\Sigma}$ having a **matrix-$F$ prior**:

    ```
    SigmaInv <- rwish(v=nu+k,S=solve(SS + Psi)
    Psi <- rwish(v=nu+delta+k-1,S=solve(SigmaInv+B0Inv))
    ```

# Properties of the matrix-$F$ distribution

- **Implementation in R:**
  - In R, draw $\mathbf{\Sigma}$ having an **inverse Wishart prior**:

    ```
    Sigma <- solve(rwish(v=n+k,S=solve(SS + B0))
    ```

  - In R, draw $\mathbf{\Sigma}$ having a **matrix-$F$ prior**:

    ```
    SigmaInv <- rwish(v=nu+k,S=solve(SS + Psi)
    Psi <- rwish(v=nu+delta+k-1,S=solve(SigmaInv+B0Inv))
    ```

- **Setting hyperparameters**
  A minimally informative default prior can be obtained by setting $\nu = k$, $\delta = 1$, and $\mathbf{B}$ equal to a "prior guess", or use an empirical Bayes prior scale (Kass & Natarajan, 2008).

# Outline

# The matrix-$F$ distribution in regularized regression

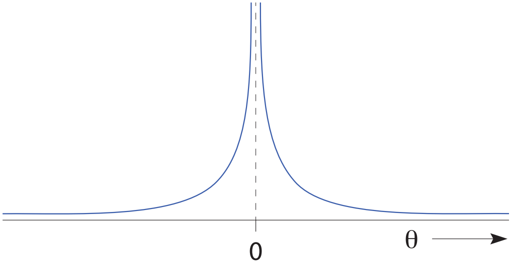
- A common problem in regression analysis is detecting true large effects in the case of many predictors ($p \gg n$).

# The matrix-$F$ distribution in regularized regression

- A common problem in regression analysis is detecting true large effects in the case of many predictors ($p \gg n$).
- The **lasso** estimate is a popular solution for this problem.

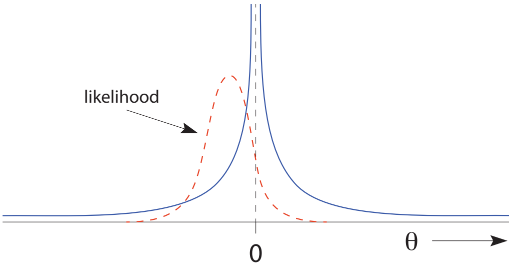
# The matrix-$F$ distribution in regularized regression

- A common problem in regression analysis is detecting true large effects in the case of many predictors ($p \gg n$).
- The **lasso** estimate is a popular solution for this problem.
- A proper **horseshoe prior** for Bayesian regularized regression performs better in certain scenario's (Carvalho et al., 2010).



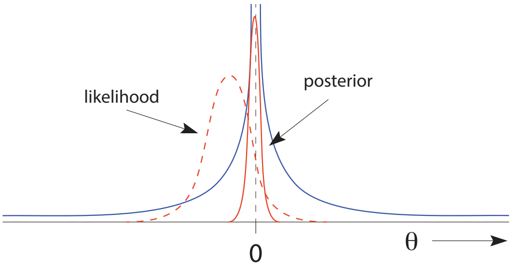
# The matrix-$F$ distribution in regularized regression

- A common problem in regression analysis is detecting true large effects in the case of many predictors ($p \gg n$).
- The **lasso** estimate is a popular solution for this problem.
- A proper **horseshoe prior** for Bayesian regularized regression performs better in certain scenario's (Carvalho et al., 2010).



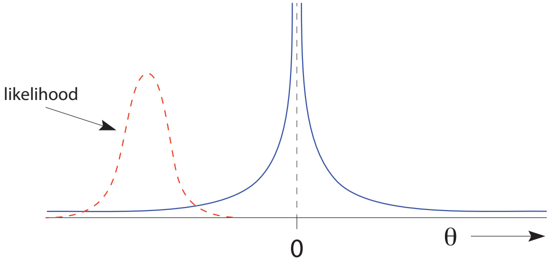
# The matrix-$F$ distribution in regularized regression

- A common problem in regression analysis is detecting true large effects in the case of many predictors ($p \gg n$).
- The **lasso** estimate is a popular solution for this problem.
- A proper **horseshoe prior** for Bayesian regularized regression performs better in certain scenario's (Carvalho et al., 2010).



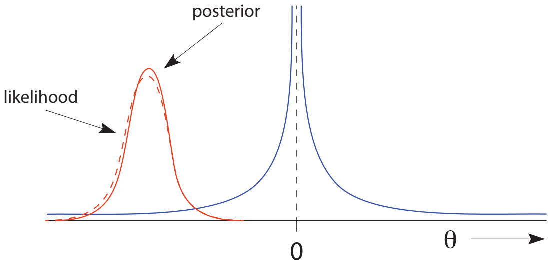
# The matrix-$F$ distribution in regularized regression

- A common problem in regression analysis is detecting true large effects in the case of many predictors ($p \gg n$).
- The **lasso** estimate is a popular solution for this problem.
- A proper **horseshoe prior** for Bayesian regularized regression performs better in certain scenario's (Carvalho et al., 2010).



# The matrix-$F$ distribution in regularized regression

- A common problem in regression analysis is detecting true large effects in the case of many predictors ($p \gg n$).
- The **lasso** estimate is a popular solution for this problem.
- A proper **horseshoe prior** for Bayesian regularized regression performs better in certain scenario's (Carvalho et al., 2010).



# The matrix-$F$ distribution in regularized regression

- When predictors are grouped, e.g., when using several dummy variables to model a categorical predictor, it may be preferable to either select all predictors belonging to a certain group or none.

# The matrix-$F$ distribution in regularized regression

- When predictors are grouped, e.g., when using several dummy variables to model a categorical predictor, it may be preferable to either select all predictors belonging to a certain group or none.
- The **grouped-lasso** is a popular solution for such grouped predictors.

# The matrix-$F$ distribution in regularized regression

- When predictors are grouped, e.g., when using several dummy variables to model a categorical predictor, it may be preferable to either select all predictors belonging to a certain group or none.
- The **grouped-lasso** is a popular solution for such grouped predictors.
- A **horse-shoe type prior** can be constructed using the matrix-$F$ distribution resulting in similar selection beheavior:

$$p(\boldsymbol{\theta}) = \int \mathcal{N}(\boldsymbol{\theta}; \mathbf{0}, \boldsymbol{\Sigma}) \times \mathcal{F}(\boldsymbol{\Sigma}; k, 1, \mathbf{B}) d\boldsymbol{\Sigma}$$

# The matrix-$F$ distribution in regularized regression

- When predictors are grouped, e.g., when using several dummy variables to model a categorical predictor, it may be preferable to either select all predictors belonging to a certain group or none.
- The **grouped-lasso** is a popular solution for such grouped predictors.
- A **horse-shoe type prior** can be constructed using the matrix-$F$ distribution resulting in similar selection beheavior:

$$p(\boldsymbol{\theta}) \quad = \quad \int \mathcal{N}(\boldsymbol{\theta}; \mathbf{0}, \boldsymbol{\Sigma}) \times \mathcal{F}(\boldsymbol{\Sigma}; k, 1, \mathbf{B}) d\boldsymbol{\Sigma}$$

  - Thicker tails than a Cauchy distribution:

$$p(\boldsymbol{\theta}) = \int \mathcal{C}(\boldsymbol{\theta}; \mathbf{0}, \boldsymbol{\Psi}) \times \mathcal{W}(\boldsymbol{\Psi}; k, \mathbf{B}) d\boldsymbol{\Psi}$$

# The matrix-$F$ distribution in regularized regression

- When predictors are grouped, e.g., when using several dummy variables to model a categorical predictor, it may be preferable to either select all predictors belonging to a certain group or none.
- The **grouped-lasso** is a popular solution for such grouped predictors.
- A **horse-shoe type prior** can be constructed using the matrix-$F$ distribution resulting in similar selection beheavior:

$$p(\boldsymbol{\theta}) \quad = \quad \int \mathcal{N}(\boldsymbol{\theta}; \mathbf{0}, \boldsymbol{\Sigma}) \times \mathcal{F}(\boldsymbol{\Sigma}; k, 1, \mathbf{B}) d\boldsymbol{\Sigma}$$

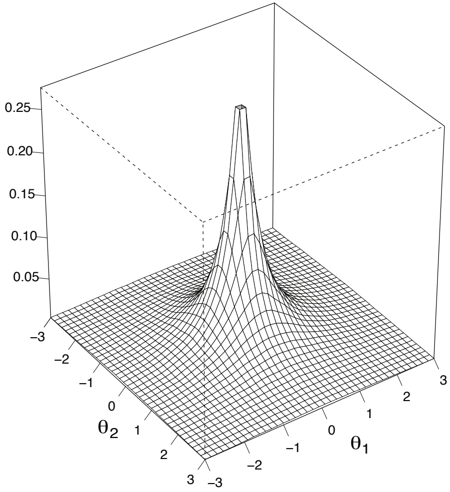
  - Thicker tails than a Cauchy distribution:

$$p(\boldsymbol{\theta}) = \int \mathcal{C}(\boldsymbol{\theta}; \mathbf{0}, \boldsymbol{\Psi}) \times \mathcal{W}(\boldsymbol{\Psi}; k, \mathbf{B}) d\boldsymbol{\Psi}$$

  - Pole at $\boldsymbol{\theta} = \mathbf{0}$ because

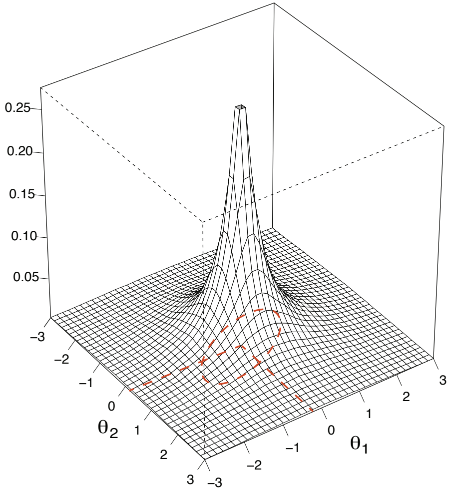
$$p(\boldsymbol{\theta}) \rightarrow +\infty \text{ as } \boldsymbol{\theta} \rightarrow \mathbf{0}.$$

# The matrix-$F$ distribution in regularized regression



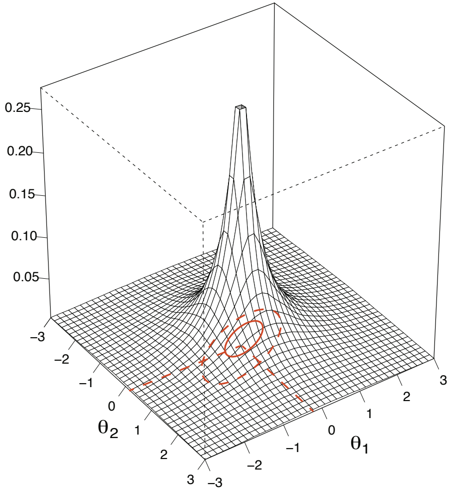

Legend: Dashed contour = likelihood contour; solid contour = posterior contour.

# The matrix-$F$ distribution in regularized regression



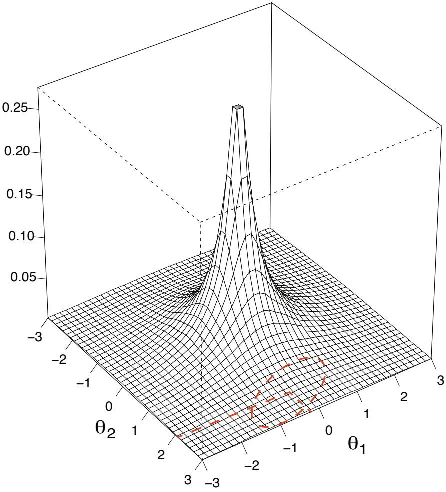

Legend: Dashed contour = likelihood contour; solid contour = posterior contour.

# The matrix-$F$ distribution in regularized regression



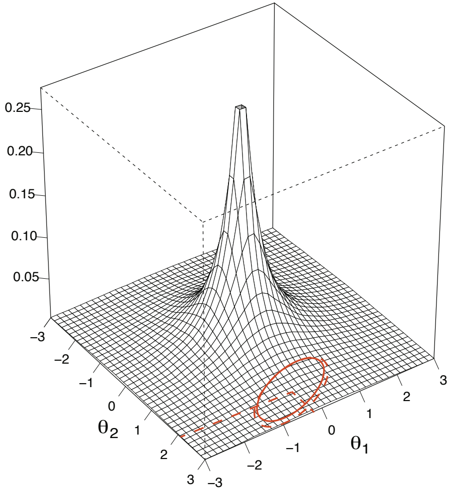

Legend: Dashed contour = likelihood contour; solid contour = posterior contour.

# The matrix-$F$ distribution in regularized regression




Legend: Dashed contour = likelihood contour; solid contour = posterior contour.

# The matrix-$F$ distribution in regularized regression




Legend: Dashed contour = likelihood contour; solid contour = posterior contour.

# Outline

# The matrix-$F$ prior for testing covariance matrices (1)

- Consider the following hypothesis test of a covariance matrix:

$$H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \text{ vs } H_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0,$$

when considering multivariate normal data, $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

# The matrix-$F$ prior for testing covariance matrices (1)

- Consider the following hypothesis test of a covariance matrix:

$$H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \text{ vs } H_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0,$$

when considering multivariate normal data, $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- Bayesian hypothesis tests can be conducted using the marginal likelihood:

$$\begin{aligned} m_0(\mathbf{X}) &= \int p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_0) p_0(\boldsymbol{\mu}) d\boldsymbol{\mu} \\ m_1(\mathbf{X}) &= \int p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) p_1(\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{\mu} d\boldsymbol{\Sigma}. \end{aligned}$$

The test is performed using the Bayes factor: $B_{01} = \frac{m_0(\mathbf{X})}{m_1(\mathbf{X})}$.

# The matrix-$F$ prior for testing covariance matrices (1)

- Consider the following hypothesis test of a covariance matrix:

$$H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \text{ vs } H_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0,$$

  when considering multivariate normal data, $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- Bayesian hypothesis tests can be conducted using the marginal likelihood:

$$\begin{aligned} m_0(\mathbf{X}) &= \int p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)p_0(\boldsymbol{\mu})d\boldsymbol{\mu} \\ m_1(\mathbf{X}) &= \int p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma})p_1(\boldsymbol{\mu}, \boldsymbol{\Sigma})d\boldsymbol{\mu}d\boldsymbol{\Sigma}. \end{aligned}$$

  The test is performed using the Bayes factor: $B_{01} = \frac{m_0(\mathbf{X})}{m_1(\mathbf{X})}$.

- **Problem**: How to choose the priors $p_0$ and $p_1$?

# The matrix-$F$ prior for testing covariance matrices (1)

- Default Bayes factors, such as the intrinsic Bayes factor (Berger & Pericchi, 1996) or the fractional Bayes factor (O'Hagan, 1995), avoid the choice of a prior by updating a noninformative improper prior with a minimal subset of the data to obtain a posterior prior, and the remaining subset of the data is used for hypothesis testing.

# The matrix-$F$ prior for testing covariance matrices (1)

- Default Bayes factors, such as the intrinsic Bayes factor (Berger & Pericchi, 1996) or the fractional Bayes factor (O'Hagan, 1995), avoid the choice of a prior by updating a noninformative improper prior with a minimal subset of the data to obtain a posterior prior, and the remaining subset of the data is used for hypothesis testing.
- In certain situations, such default Bayes factors behave as actual Bayes factors based on so-called intrinsic priors as $n \to \infty$.

# The matrix-$F$ prior for testing covariance matrices (1)

- Default Bayes factors, such as the intrinsic Bayes factor (Berger & Pericchi, 1996) or the fractional Bayes factor (O'Hagan, 1995), avoid the choice of a prior by updating a noninformative improper prior with a minimal subset of the data to obtain a posterior prior, and the remaining subset of the data is used for hypothesis testing.
- In certain situations, such default Bayes factors behave as actual Bayes factors based on so-called intrinsic priors as $n \to \infty$.
- A proper intrinsic prior can be used to compute an "objective" Bayes factor without needing to formulate a subjective prior or without needing to split the data for prior specification and hypothesis testing.

# The matrix-$F$ prior for testing covariance matrices (1)

- An intrinsic prior can be found via (Berger & Pericchi, 2004)

$$p_1^N(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{X}_{(\ell)})$$

# The matrix-$F$ prior for testing covariance matrices (1)

- An intrinsic prior can be found via (Berger & Pericchi, 2004)

$$p_1^N(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}_{(\ell)})p_0^N(\mathbf{X}_{(\ell)})$$

# The matrix-$F$ prior for testing covariance matrices (1)

- An intrinsic prior can be found via (Berger & Pericchi, 2004)

$$p_1^I(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int p_1^N(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}_{(\ell)})p_0^N(\mathbf{X}_{(\ell)})d\mathbf{X}_{(\ell)}$$

# The matrix-$F$ prior for testing covariance matrices (1)

- An intrinsic prior can be found via (Berger & Pericchi, 2004)

$$p_1^I(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \int p_1^N(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{X}_{(\ell)})p_0^N(\mathbf{X}_{(\ell)})d\mathbf{X}_{(\ell)},$$

where

$$\begin{aligned} p_0^N(\mathbf{X}_{(\ell)}) &= \int p(\mathbf{X}_{(\ell)}|\boldsymbol{\mu},\boldsymbol{\Sigma}_0)p_0^N(\boldsymbol{\mu})d\boldsymbol{\mu} \\ p_1^N(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{X}_{(\ell)}) &= \frac{p(\mathbf{X}_{(\ell)}|\boldsymbol{\mu},\boldsymbol{\Sigma})p_1^N(\boldsymbol{\mu},\boldsymbol{\Sigma})}{\int p(\mathbf{X}_{(\ell)}|\boldsymbol{\mu},\boldsymbol{\Sigma})p_1^N(\boldsymbol{\mu},\boldsymbol{\Sigma})d\boldsymbol{\mu}d\boldsymbol{\Sigma}} \end{aligned}$$

# The matrix-$F$ prior for testing covariance matrices (1)

### Theorem

When testing $H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}_0$ versus $H_1 : \mathbf{\Sigma} \neq \mathbf{\Sigma}_0$ using iid $k$-variate data with $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$, for $i = 1, \ldots, n$, the intrinsic prior under $H_1$ is given by

$$\pi_1^I(\boldsymbol{\mu}, \mathbf{\Sigma}) = F(\mathbf{\Sigma}; k, 1, \mathbf{\Sigma}_0)$$

based on the noninformative improper priors $\pi_1^N(\boldsymbol{\mu}, \mathbf{\Sigma}) = |\mathbf{\Sigma}|^{-\frac{k+1}{2}}$ and $\pi_0^N(\boldsymbol{\mu}) = 1$, and a minimal training sample of size $m = k + 1$. This is also the case when $\boldsymbol{\mu}$ is known.

### Proposition

The Bayes factor of $H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}_0$ versus $H_1 : \mathbf{\Sigma} \neq \mathbf{\Sigma}_0$ based on the intrinsic prior is consistent.

# The matrix-$F$ prior for testing covariance matrices (2)

- Consider the following hypothesis test of a covariance matrix:

  $H_1 : \sigma_1 < \ldots < \sigma_k$ vs $H_2 : \sigma_1 > \ldots > \sigma_k$ vs $H_3$ : neither $H_1$, nor $H_2$,

  when considering multivariate normal data, $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

# The matrix-$F$ prior for testing covariance matrices (2)

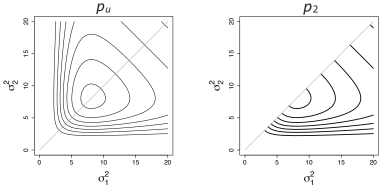
- Consider the following hypothesis test of a covariance matrix:

$$H_1 : \sigma_1 < \ldots < \sigma_k \text{ vs } H_2 : \sigma_1 > \ldots > \sigma_k \text{ vs } H_3 : \text{neither } H_1, \text{ nor } H_2,$$

when considering multivariate normal data, $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- Let $H_u$ : "$\boldsymbol{\Sigma}$ is pos. def.", and

$$p_2(\boldsymbol{\Sigma}) = p_u(\boldsymbol{\Sigma}) \times I(\sigma_1 > \ldots > \sigma_k) \times Pr(\sigma_1 > \ldots > \sigma_k | H_u)^{-1}$$



# The matrix-$F$ prior for testing covariance matrices (2)

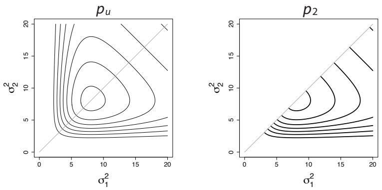
- Consider the following hypothesis test of a covariance matrix:

  $$H_1 : \sigma_1 < \ldots < \sigma_k \text{ vs } H_2 : \sigma_1 > \ldots > \sigma_k \text{ vs } H_3 : \text{neither } H_1, \text{ nor } H_2,$$

  when considering multivariate normal data, $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- Let $H_u$ : "$\boldsymbol{\Sigma}$ is pos. def.", and

  $$p_2(\boldsymbol{\Sigma}) = p_u(\boldsymbol{\Sigma}) \times I(\sigma_1 > \ldots > \sigma_k) \times Pr(\sigma_1 > \ldots > \sigma_k | H_u)^{-1}$$




- The Bayes factor is given by: $B_{2u} = \frac{Pr(\sigma_1 > \ldots > \sigma_k | H_u, \mathbf{X})}{Pr(\sigma_1 > \ldots > \sigma_k | H_u)}$.

# The matrix-$F$ prior for testing covariance matrices (2)

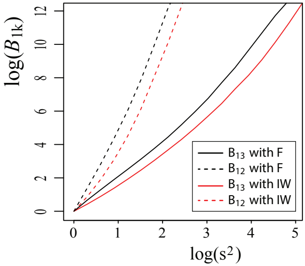
- $H_1 : \sigma_1 < \ldots < \sigma_k$ vs $H_2 : \sigma_1 > \ldots > \sigma_k$ vs $H_3$ : neither $H_1$, nor $H_2$.

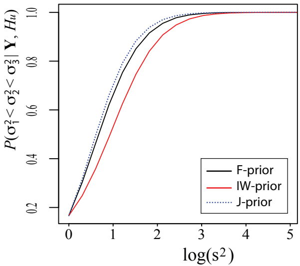
# The matrix-$F$ prior for testing covariance matrices (2)

- $H_1 : \sigma_1 < \ldots < \sigma_k$ vs $H_2 : \sigma_1 > \ldots > \sigma_k$ vs $H_3$ : neither $H_1$, nor $H_2$.
- As unconstrained priors we considered
  1. $\boldsymbol{\Sigma} \sim \mathcal{F}(3, 1, \mathbf{I}_3)$.
  2. $\boldsymbol{\Sigma} \sim \mathcal{IW}(3, \mathbf{I}_3)$.

# The matrix-$F$ prior for testing covariance matrices (2)

- $H_1 : \sigma_1 < \ldots < \sigma_k$ vs $H_2 : \sigma_1 > \ldots > \sigma_k$ vs $H_3$ : neither $H_1$, nor $H_2$.
- As unconstrained priors we considered
  1. $\boldsymbol{\Sigma} \sim \mathcal{F}(3, 1, \mathbf{I}_3)$.
  2. $\boldsymbol{\Sigma} \sim \mathcal{IW}(3, \mathbf{I}_3)$.
- We fixed $n = 20$ and let $\mathbf{S} = diag(1, s, s^2)$, while $s \to \infty$.



# Outline

# The matrix-$F$ prior for estimating hierarchical models (1)

- Kass and Natarajan (2006) considered the following hierarchical Poisson regression model:

$$\begin{aligned} y_i | b_i, x_i &\sim Poisson(\mu_i^{x,b}) \\ \mu_I^{x,b} &= \exp\{\beta_0 + \beta_1 \log(x_i + 10) + \beta_2 x_i + b_i\} \\ b_i &\sim N(0, \sigma^2), \end{aligned}$$

for $i = 1, \ldots, 18$.

# The matrix-$F$ prior for estimating hierarchical models (1)

- Kass and Natarajan (2006) considered the following hierarchical Poisson regression model:

$$\begin{aligned} y_i | b_i, x_i &\sim Poisson(\mu_i^{x,b}) \\ \mu_I^{x,b} &= \exp\{\beta_0 + \beta_1 \log(x_i + 10) + \beta_2 x_i + b_i\} \\ b_i &\sim N(0, \sigma^2), \end{aligned}$$

  for $i = 1, \ldots, 18$.

- Population values: $\beta_0 = 2.203$, $\beta_1 = .311$, $\beta_2 = -.001$, and $\sigma^2 = .04$.

# The matrix-$F$ prior for estimating hierarchical models (1)

- Kass and Natarajan (2006) considered the following hierarchical Poisson regression model:

$$\begin{aligned} y_i|b_i, x_i &\sim Poisson(\mu_i^{x,b}) \\ \mu_I^{x,b} &= \exp\{\beta_0 + \beta_1 \log(x_i + 10) + \beta_2 x_i + b_i\} \\ b_i &\sim N(0, \sigma^2), \end{aligned}$$

  for $i = 1, \ldots, 18$.
- Population values: $\beta_0 = 2.203$, $\beta_1 = .311$, $\beta_2 = -.001$, and $\sigma^2 = .04$.
- Classical risk and nonconvergence of the 95%-CI's were determined.

# The matrix-$F$ prior for estimating hierarchical models (1)

## Hierarchical Poisson regression model

| | $IW(1,R^*)$ | $\pi_{us}$ | $F(1,1,R^*)$ | $F(1,1,10^3)$ | $(\sigma^2)^{-\frac{1}{2}}$ |
|---|---|---|---|---|---|
| Risk | | | | | |
| $\boldsymbol{\beta}$ | $.01 \pm .00$ | $.01 \pm .00$ | $.11 \pm .00$ | $.10 \pm .00$ | $.11 \pm .00$ |
| $\sigma^2$ | $.12 \pm .00$ | $.62 \pm .02$ | $.23 \pm .01$ | $.28 \pm .01$ | $.27 \pm .01$ |
| Noncoverage | | | | | |
| $\beta_0$ | $.056 \pm .007$ | $.070 \pm .008$ | $.064 \pm .007$ | $.047 \pm .007$ | $.048 \pm .008$ |
| $\beta_1$ | $.059 \pm .007$ | $.067 \pm .008$ | $.065 \pm .007$ | $.048 \pm .007$ | $.049 \pm .007$ |
| $\beta_2$ | $.060 \pm .007$ | $.075 \pm .008$ | $.053 \pm .007$ | $.058 \pm .007$ | $.051 \pm .007$ |
| $\sigma^2$ | $.007 \pm .003$ | $.037 \pm .006$ | $.048 \pm .007$ | $.050 \pm .007$ | $.045 \pm .007$ |

- $IW(1,R^*)$ is the default (empirical Bayes) conjugate prior of Kass & Natarajan (2006);
- $\pi_{us}$ is the approximate uniform shrinkage prior of Natarajan & Kass (1999).

# The matrix-$F$ prior for estimating hierarchical models (2)

- Natarajan and Kass (1999) considered the following hierarchical logistic regression model:

$$
\begin{aligned}
\text{logit}(\mu_{ij}^{\mathbf{b}}) &= \beta_0 + \beta_1 t_j + \beta_2 x_i + \beta_3 x_i t_j + b_{i0} + b_{i1} t_j \\
\mathbf{b}_i &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),
\end{aligned}
$$

for $n = 30$, $t_j = j - 4$, for $j = 1, \ldots, 7$.

# The matrix-$F$ prior for estimating hierarchical models (2)

- Natarajan and Kass (1999) considered the following hierarchical logistic regression model:

$$\begin{aligned} \text{logit}(\mu_{ij}^{\mathbf{b}}) &= \beta_0 + \beta_1 t_j + \beta_2 x_i + \beta_3 x_i t_j + b_{i0} + b_{i1} t_j \\ \mathbf{b}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \end{aligned}$$

  for $n = 30$, $t_j = j - 4$, for $j = 1, \ldots, 7$.

- Population values: $\boldsymbol{\beta} = (-.625, .25, -.25, .125)'$ and $\mathbf{\Sigma} = \text{diag}(.5, .25)$.

# The matrix-$F$ prior for estimating hierarchical models (2)

- Natarajan and Kass (1999) considered the following hierarchical logistic regression model:

$$\begin{aligned} \text{logit}(\mu_{ij}^{\mathbf{b}}) &= \beta_0 + \beta_1 t_j + \beta_2 x_i + \beta_3 x_i t_j + b_{i0} + b_{i1} t_j \\ \mathbf{b}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \end{aligned}$$

  for $n = 30$, $t_j = j - 4$, for $j = 1, \ldots, 7$.
- Population values: $\boldsymbol{\beta} = (-.625, .25, -.25, .125)'$ and $\mathbf{\Sigma} = \text{diag}(.5, .25)$.
- Classical risk and nonconvergence of the 95%-CI's were determined.

# The matrix-$F$ prior for estimating hierarchical models (2)

**Hierarchical logistic regression model**

Results for the random effects covariance matrix $\boldsymbol{\Sigma}$.

| | | Noncoverage | | | Interval width | | |
|---|---|---|---|---|---|---|---|
| Prior | Risk | $\sigma_1^2$ | $\sigma_{12}$ | $\sigma_2^2$ | $\sigma_1^2$ | $\sigma_{12}$ | $\sigma_2^2$ |
| $F(\boldsymbol{\Sigma}; 2, 2, \mathbf{R}^*)$ | $3.32 \pm .18$ | .034 | .045 | .043 | 2.11 | 1.07 | .90 |
| $\pi_{us}$ | $3.10 \pm .19$ | .035 | .029 | .041 | 2.12 | 1.05 | .88 |
| HW-prior | $7.64 \pm .50$ | .070 | .009 | .110 | 2.89 | 1.08 | 1.28 |

- $\pi_{us}$ is the approximate uniform shrinkage prior of Natarajan & Kass (1999).
- The HW-prior is the marginally noninformative prior of Huang and Wand (2013).

# The matrix-$F$ prior for estimating hierarchical models (2)

**Hierarchical logistic regression model**

Results for the fixed effects $\boldsymbol{\beta}$.

| | | Noncoverage | | | | Interval width | | | |
|---|---|---|---|---|---|---|---|---|---|
| Prior | Risk | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
| $F(\boldsymbol{\Sigma}; 2, 2, \mathbf{R}^*)$ | $.44 \pm .01$ | .052 | .048 | .055 | .045 | 1.33 | .81 | 1.89 | 1.15 |
| $\pi_{us}$ | $.46 \pm .02$ | .033 | .058 | .044 | .045 | 1.44 | .83 | 2.12 | 1.19 |
| HW-prior | $.51 \pm .02$ | .061 | .046 | .055 | .044 | 1.45 | .91 | 2.05 | 1.28 |

- $\pi_{us}$ is the approximate uniform shrinkage prior of Natarajan & Kass (1999).
- The HW-prior is the marginally noninformative prior of Huang and Wand (2013).

# The matrix-$F$ prior for estimating hierarchical models (2)

**Hierarchical logistic regression model**

Results for the random effects $\mathbf{b}_i$.

| | Risk | | Noncoverage | | Interval width | |
|---|---|---|---|---|---|---|
| Prior | $b_0$ | $b_1$ | $b_0$ | $b_1$ | $b_0$ | $b_1$ |
| $F(\mathbf{\Sigma}; 2, 2, \mathbf{R}^*)$ | $11.65 \pm .13$ | $4.67 \pm .05$ | .058 | .057 | 2.54 | 1.60 |
| $\pi_{us}$ | $11.51 \pm .12$ | $4.51 \pm .05$ | .045 | .048 | 2.67 | 1.63 |
| HW-prior | $12.46 \pm .17$ | $5.20 \pm .08$ | .049 | .046 | 2.80 | 1.77 |

- $\pi_{us}$ is the approximate uniform shrinkage prior of Natarajan & Kass (1999).
- The HW-prior is the marginally noninformative prior of Huang and Wand (2013).

# Outline

# Summary

- The $F$ distribution can "safely" be used as prior for the random effects covariance matrix.
- The matrix-$F$ prior is competitive in terms of risk and coverage rates in generalized linear mixed models.
- The matrix-$F$ prior can straightforwardly be implemented in a Gibbs sampler.
- A minimally informative matrix-$F$ prior can easily be specified based on a prior guess or empirical Bayes scale matrix.
- The matrix-$F$ prior can be used for constructing multivariate horseshoe type priors for estimating sparse signals.
- The matrix-$F$ prior serves as an intrinsic prior when testing a covariance matrix of multivariate normal data.
- The matrix-$F$ prior results in satisfactory selection behavior for testing inequality constrained hypotheses.

# References


- Berger, J. O. & Pericchi, L. R. (2004). Training samples in objective Bayesian model selection. *The Annals of Statistics, 32*, 841–869.
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika, 97*, 465–480.
- Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika, 68*, 265–274.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models(Comment on Article by Browne and Draper). *Bayesian Analysis, 3*, 515–534.
- Mulder, J. & Pericchi, L.R. (in press). The matrix-variate F prior for estimating and testing covariance matrices. *Bayesian Analysis*.

Thank you!