Graphical Evidence

Anindya Bhadra
www.stat.purdue.edu/~bhadra

Purdue University
Overview

- Marginal likelihood or *evidence* is fundamental to Bayesian statistics.

- Used for empirical Bayes tuning of hyperparameters, model selection using Bayes factors.

- There is no dearth of generic approaches, yet calculation of evidence is mostly unresolved in Gaussian graphical models (GGMs), except for very specific priors such as the Wishart or G-Wishart.

- **Goal**: To provide a tractable approach for evidence calculation in GGMs under mild requirements.

- **Joint work with Ksheera Sagar (Purdue), Sayantan Banerjee (IIM Indore) and Jyotishka Datta (Virginia Tech).**
Suppose $y_{n \times p} \sim \mathcal{N}(0, I_n \otimes \Omega_{p \times p}^{-1})$. Evidence calculation is simple in principle:

$$f(y) = \int_{\Omega \in \mathcal{M}_p^+} f(y | \Omega) f(\Omega) d\Omega.$$ 

The restriction of the integral to the space of positive definite matrices causes a lot of difficulties, except for Wishart and specific instances of G-Wishart (Uhler et al., 2018, AoS).

For the same reason, a “default” covering density is very hard to design: difficulties for importance, bridge or path sampling.
Generic approaches for estimating evidence

- Harmonic mean estimates and variants (Newton and Raftery, 1994, JRSSB; Gelfand and Dey, 1994, JRSSB)

- Importance sampling approaches:
  - Annealed importance sampling (Neal, 2001, Stats. Comput.)

- Nested sampling (Skilling, 2006, BA).

- Chib (1995, JASA) and Chib and Jeliazkov (2001, JASA) based on MCMC posterior draws.

- Excellent review article by Llorente et al. (2022, SIAM Review).
Do generic approaches work in GGMs?

- HM estimates can have unbounded variance: limit distribution is $\alpha$ stable (Wolpert and Schmeider, 2012).

- We are not aware of any principled way of choosing an importance or bridge density under a positive definite restriction.

- Nested sampling requires sampling from a progressively higher likelihood region: very hard to implement in high dimensions.

- **A case in point:** the specialized Monte Carlo method of Atay-Kayis and Massam (2005, Biometrika) for G-Wishart marginals appeared a good 10 years after these generic approaches.
Chib (1995)

- Recall the fundamental Bayesian identity:

\[ f(y) = \frac{f(y | \theta)f(\theta)}{f(\theta | y)}, \]

- The likelihood and the prior can typically be evaluated at some \( \theta = \theta^* \), the trouble is evaluating \( f(\theta | y) \).

- Chib’s strategy:
  - Decompose \( \Omega = (z, \theta) = (\text{nuisance parameter}, \text{parameter of interest}) \).
  - Run a Gibbs sampler iterating between \( f(z | \theta, y) \) and \( f(\theta | z, y) \). Converges to \( f(z, \theta | y) \). Correct marginals for \((z | y)\) and \((\theta | y)\).
  - Estimate using the Gibbs draws:

\[ \hat{f}(\theta^* | y) = M^{-1}\sum_{i=1}^{M} f(\theta^* | z^{(i)}, y), \quad z^{(i)} \sim f(z | y). \]

- Need the constants only for \( f(\theta | z, y) \); not for \( f(z | \theta, y) \).
Pros and cons of Chib (1995)

- Chib’s approach is automatic in the same way a Gibbs sampler is automatic: a covering (importance, bridge) density is not required.

- But applying Chib’s method requires designing a suitable $f(\theta \mid z, y)$ that can be evaluated (merely sampling from it is not enough).

- Application is a matter of art and not generic in a way the harmonic mean estimate is generic.

- Some known difficulties in finite mixture models (Neal, 1999).
Chib’s approach for GGMs: the telescoping block decomposition

- Apply the decomposition:
  \[
  \Omega_{p \times p} = \begin{bmatrix}
  \Omega_{(p-1) \times (p-1)} & \omega_{\cdot p} \\
  \omega_{\cdot p}^T & \omega_{pp}
  \end{bmatrix}.
  \]

- Let \( \theta_p = (\omega_{\cdot p}, \omega_{pp}) \) and \( z = \) collection of all other latent variables.

- Wang (2012, BA) showed in the context of sampling that \( f(\theta_p | y, z) = f(\omega_{\cdot p}, \omega_{pp} | y, z) = f(\omega_{\cdot p} | y, z) f(\omega_{pp} | \omega_{\cdot p}, y, z) \) decomposes as \((\text{normal } \times \text{ gamma})\) under suitable priors on \( \Omega_{p \times p}. \)

- We will use this for density evaluation, since the normalizing constants for both normal and gamma densities are available!
Chib’s approach for GGMs: the telescoping block decomposition

- We have

\[
\log f(y_{1:p}) = \log f(y_{1:p} | \theta_p) + \log f(\theta_p) - \log f(\theta_p | y_{1:p}).
\]

- Slightly rewrite:

\[
\log f(y_{1:p}) = \log f(y_p | y_{1:p-1}, \theta_p) + \log f(y_{1:p-1} | \theta_p) + \log f(\theta_p) - \log f(\theta_p | y_{1:p}) \\
:= I_p + II_p + III_p - IV_p.
\]

- We can evaluate the partial likelihood \(I_p\) using

\[
y_p | y_{1:p-1}, \theta_p \sim N(-y_{1:p-1} \omega_{\bullet p} / \omega_{pp}, 1/\omega_{pp}),
\]

- Assume \(III_p\) can be evaluated and Wang’s result from the previous slide will be used for evaluating \(IV_p\). There remains \(II_p\) to deal with.
Chib’s approach for GGMs: the telescoping block decomposition

- BUT! The term $\Pi$ is telescoping. We have:

\[
\Pi_p = \log f(y_{1:p-1} | \theta_p) \\
= \log f(y_{p-1} | y_{1:p-2}, \theta_p, \theta_{p-1}) + \log f(y_{1:p-2} | \theta_p, \theta_{p-1}) \\
+ \log f(\theta_{p-1} | \theta_p) - \log f(\theta_{p-1} | y_{1:p-1}, \theta_p) \\
:= I_{p-1} + \Pi_{p-1} + \Pi I_{p-1} - \Pi V_{p-1}.
\]

- We use a form of iterative proportional scaling (IPS). Define $\tilde{\Omega}(p-1) \times (p-1)$ as:

\[
\tilde{\Omega}(p-1) \times (p-1) = \Omega(p-1) \times (p-1) - \frac{\omega \omega^T}{\omega_p} := \begin{bmatrix} \Omega(p-2) \times (p-2) & \tilde{\omega}(p-1) \\
\tilde{\omega}(p-1)^T & \tilde{\omega}(p-1)(p-1) \end{bmatrix}.
\]

Then $\tilde{\Omega}(p-1) \times (p-1)$ is p.d. and $(y_{1:p-1} | \theta_p, \Omega(p-1) \times (p-1)) \sim \mathcal{N}(0, \tilde{\Omega}^{-1}(p-1) \times (p-1))$.

- Thus, $I_{p-1}$ can be evaluated using:

\[
y_{p-1} | y_{1:p-2}, \theta_p, \theta_{p-1} \sim \mathcal{N}(-y_{1:p-2} \tilde{\omega}(p-1), 1/\tilde{\omega}(p-1)(p-1)).
\]
Overall strategy

Figure: (a) Decomposition of \( \Omega_{p \times p} \). Purple, green and blue blocks denote \( \theta_p, \theta_{p-1} \) and finally \( \theta_1 = \omega_{11} \). Red arrow denotes how the algorithm proceeds, fixing one row/column at a time, and (b) the telescoping sum giving the log-marginal log \( f(y_{1:p}) \).

- Run Chib \( p \) times, adjusting \( \Omega \) each time. In each equation, evaluate only I, III and IV. Eliminate II via telescoping sum.
A demonstration on Wishart (known marginal)

- Suppose $\Omega \sim \mathcal{W}_p(\mathbf{V}, \alpha)$. Then,

$$\log f(y_1:p) = -\frac{np}{2} \log(\pi) + \log \Gamma_p \left(\frac{\alpha + n}{2}\right) - \log \Gamma_p \left(\frac{\alpha}{2}\right) + \frac{(\alpha + n)}{2} \log \left| \mathbf{I}_p + \mathbf{V}^{1/2} \mathbf{S} \mathbf{V}^{1/2} \right|.$$ 

- The closed form expression for the marginal provides an oracle.
Computing $\text{III}_p \ (\equiv \log f(\theta_p))$

- Recall, $\theta_p = (\omega_{\cdot p}, \omega_{pp})$.

- If $\Omega \sim \mathcal{W}_p(I_p, \alpha)$ then $f(\omega_{\cdot p}, \omega_{pp}) = f(\omega_{\cdot p} | \omega_{pp})f(\omega_{pp})$, where,

  $\omega_{\cdot p} | \omega_{pp} \sim \mathcal{N}(0, \omega_{pp}I_{p-1}), \omega_{pp} \sim \text{Gamma}(\text{shape} = \alpha/2, \text{rate} = 1/2)$.

- Computing $\text{III}_p$ is easy: normal $\times$ gamma.
Computing $\text{IV}_p (\equiv \log f(\theta_p \mid y_{1:p}))$

- Decompose $S = y^T y$ analogous to $\Omega$ and reparameterize $(\omega_{\bullet p}, \omega_{pp}) \mapsto (\beta_{\bullet p}, \gamma_{pp})$:

$$S = \begin{bmatrix} S_{(p-1) \times (p-1)} & s_{\bullet p} \\ s_{\bullet p}^T & s_{pp} \end{bmatrix}, \quad \beta_{\bullet p} = \omega_{\bullet p}, \quad \gamma_{pp} = \omega_{pp} - \omega_{\bullet p}^T \Omega_{(p-1) \times (p-1)}^{-1} \omega_{\bullet p}. $$

- Key result of Wang (2012, BA):

$$f(\beta_{\bullet p}, \gamma_{pp} \mid \text{rest}) = \mathcal{N}(\beta_{\bullet p} \mid -Cs_{\bullet p}, C) \times G\left(\gamma_{pp} \mid \frac{n + \alpha - p - 1}{2} + 1, \frac{s_{pp} + 1}{2}\right).$$

where $C = \left\{(s_{pp} + 1)\Omega_{(p-1) \times (p-1)}^{-1}\right\}^{-1}$

- Allows computation of $\text{IV}_p$ using Chib's two block strategy.
Computing $\mathbf{III}_{p-1}, \ldots, \mathbf{III}_1$ and $\mathbf{IV}_{p-1}, \ldots, \mathbf{IV}_1$

- Same strategy as in going from $\mathbf{I}_p$ to $\mathbf{I}_{p-1}$.

- Proceed backwards starting from the $p$th row. At each step, adjust the upper left sub-matrix $\Omega_{j \times j}$ via IPS:

\[
\text{for } (j = p-1, \ldots, 1) \text{ do }
\]

\[
\text{Update } \Omega_{j \times j} \leftarrow \Omega_{j \times j} - \frac{\omega_{j+1}(j+1) \omega_{j+1}^T(j+1)}{\omega_{j+1,j+1}}.
\]

end for

- Calculate $\mathbf{III}_j$ and $\mathbf{IV}_j$ with the updated $\Omega_{j \times j}$. 
## Results for Wishart

<table>
<thead>
<tr>
<th>Dimension and Parameters</th>
<th>Truth</th>
<th>Proposed</th>
<th>AIS</th>
<th>Nested</th>
</tr>
</thead>
<tbody>
<tr>
<td>($p = 5$, $n = 10$, $\alpha = 7$)</td>
<td>-84.13</td>
<td>-84.13 (-0.02)</td>
<td>-84.3 (0.68)</td>
<td>-84.26 (0.57)</td>
</tr>
<tr>
<td>($p = 10$, $n = 20$, $\alpha = 13$)</td>
<td>-365.11</td>
<td>-365.12 (0.06)</td>
<td>-397.64 (6.1)</td>
<td>-392.2 (6.04)</td>
</tr>
<tr>
<td>($p = 15$, $n = 30$, $\alpha = 20$)</td>
<td>-837.7</td>
<td>-837.67 (0.13)</td>
<td>-1000.45 (13.5)</td>
<td>-994.87 (13.7)</td>
</tr>
<tr>
<td>($p = 25$, $n = 50$, $\alpha = 33$)</td>
<td>-2417.65</td>
<td>-2417.14 (1.11)</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
</tr>
<tr>
<td>($p = 30$, $n = 60$, $\alpha = 39$)</td>
<td>-3553.62</td>
<td>-3548.02 (3.04)</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
</tr>
</tbody>
</table>

**Table:** Mean (sd) of estimated log marginal for Wishart for the proposed approach, AIS, nested sampling; under 25 random permutations of the nodes \{1,\ldots,p\} using 5000 samples.
Evidence under element-wise priors

- Clearly, we did not get into all this trouble just for Wishart!

- Consider the element-wise prior:

\[
f(\Omega \mid \lambda) = C^{-1} \prod_{i<j} f(\omega_{ij} \mid \lambda) \prod_{j=1}^{p} f(\omega_{jj} \mid \lambda) \mathbb{1}(\Omega \in \mathcal{M}_p^+).
\]

- Two examples (global-local shrinkage priors):
  - Bayesian graphical lasso (BGL):
    \[
    f(\omega_{ij} \mid \lambda) = \left(\frac{\lambda}{2}\right) \exp(-\lambda |\omega_{ij}|)
    \]
    \[
    \uparrow (Andrews and Mallows, 1974)
    \]
    \[
    \omega_{ij} \mid \tau_{ij}, \lambda \sim \mathcal{N}(0, \tau_{ij}), \quad \tau_{ij} \mid \lambda \sim \text{Exp}(\lambda^2/2).
    \]
  - Graphical horseshoe (GHS):
    \[
    \omega_{ij} \mid \tau_{ij}, \lambda \sim \mathcal{N}(0, \tau_{ij}), \quad \tau_{ij} \mid \lambda \sim \mathcal{C}^+(0, \lambda).
    \]
Evidence under element-wise priors

- The off-diagonal $\omega_{ij}$ terms are normal conditional on $\tau_{ij}$.

- Similarly, the diagonal $\omega_{jj}$ are exponential.

- The presence of these mixing $\tau_{ij}$ variables is the ONLY difference with the Wishart case for our purposes.

- The $\tau_{ij}$ terms can be sampled easily.

- MAIN IDEA: Absorb the $\tau_{ij}$ terms into Chib’s latent $z$ (they are sampled, but their densities are not evaluated). Conditional on these, evaluate the normal and gamma densities exactly as in Wishart.
Computing $\text{IV}_p (\equiv \log f(\theta_p \mid y_{1:p}))$

- We have

$$f(\beta_\cdot p, \gamma_{pp} \mid \tau_\cdot p, \Omega_{(p-1)\times(p-1)}, y_{1:p}) = \mathcal{N}(\beta_\cdot p \mid -Cs_{\cdot p}, C)$$

$$\times \text{Gamma} \left( \frac{\gamma_{pp}}{2} + 1, \frac{s_{pp} + \lambda}{2} \right),$$

where $C = \left\{ \text{diag}^{-1}(\tau_\cdot p) + (s_{pp} + \lambda)\Omega_{(p-1)\times(p-1)}^{-1} \right\}^{-1}$

- Recall, for Wishart we had

$$f(\beta_\cdot p, \gamma_{pp} \mid \Omega_{(p-1)\times(p-1)}, y_{1:p}) = \mathcal{N}(\beta_\cdot p \mid -Cs_{\cdot p}, C)$$

$$\times G \left( \frac{n + \alpha - p - 1}{2} + 1, \frac{s_{pp} + 1}{2} \right).$$

where $C = \left\{ (s_{pp} + 1)\Omega_{(p-1)\times(p-1)}^{-1} \right\}^{-1}$. 
Results

Figure: Log marginal vs. $\lambda$ under (a) BGL and (b) GHS ($p = 10, n = 150$).

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>0.05</th>
<th>1</th>
<th>2 ($= \lambda_0$)</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>log BF</td>
<td>BGL</td>
<td>138.84</td>
<td>7.86</td>
<td>0.18</td>
<td>4.98</td>
<td>13.9</td>
</tr>
<tr>
<td></td>
<td>GHS</td>
<td>115.31</td>
<td>7.89</td>
<td>0.12</td>
<td>1.79</td>
<td>3.63</td>
</tr>
</tbody>
</table>

Table: Logarithm of Bayes factors.
Additional results and applications

- The strategy also works for calculating evidence under G-Wishart priors.

- Results are quite competitive with current state of the art (Atay-Kayis and Massam, 2005)

- As a by product, we are also able to develop a new row-wise sampler for G-Wishart that does not require a maximal clique decomposition.

- Details in the paper.
Concluding remarks

- The strategy developed will work whenever: (a) the priors on the off-diagonals of $\Omega$ are scale mixtures of normal and (b) the diagonals of $\Omega$ are scale mixtures of exponential.

- These are very mild requirements and can handle a broad class of priors.

- Although we did not do so in this paper, one may also shift focus from the prior to likelihood that are mixtures of normal! Consider $y \sim t_\nu(\mu, \Omega^{-1})$

- This is equivalent to $y \mid \tau \sim N(\mu, \tau^{-1}\Omega^{-1}), \tau \sim \text{Gamma}(\nu/2, \nu/2)$.

- Should be possible to absorb the $\tau$ in the likelihood into Chib’s $z$ and proceed.
Main references

  - Code: https://github.com/sagarknk/Graphical_Evidence
