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).*

Evidence in GGMs

- Suppose $\mathbf{y}_{n\times p} \sim \mathcal{N}(0, \mathbf{I}_n \otimes \mathbf{\Omega}^{-1}_{p\times p})$. Evidence calculation is simple in principle:
$$f(\mathbf{y}) = \int_{\mathbf{\Omega}\in\mathcal{M}_p^+} f(\mathbf{y} \mid \mathbf{\Omega}) f(\mathbf{\Omega}) d\mathbf{\Omega}.$$

- 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:
  - Bridge sampling and variants (path, warped bridge) (Gelman and Meng, 1998, Stats. Sci.; Meng and Wong, 1996, Sinica; Meng and Schilling, 2002, JCGS),
  - 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 α 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:

$$f(\mathbf{y}) = \frac{f(\mathbf{y} \mid \theta) f(\theta)}{f(\theta \mid \mathbf{y})},$$

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

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

$$\hat{f}(\theta^* \mid \mathbf{y}) = M^{-1} \sum_{i=1}^{M} f(\theta^* \mid z^{(i)}, \mathbf{y}), \quad z^{(i)} \sim f(z \mid \mathbf{y}).$$

  - Need the constants only for $f(\theta \mid z, \mathbf{y})$; not for $f(z \mid \theta, \mathbf{y})$.

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 $f(\theta \mid z, \mathbf{y})$ 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:
$$\boldsymbol{\Omega}_{p\times p} = \begin{bmatrix} \boldsymbol{\Omega}_{(p-1)\times(p-1)} & \boldsymbol{\omega}_{\bullet p} \\ \boldsymbol{\omega}_{\bullet p}^{T} & \omega_{pp} \end{bmatrix}.$$

- Let $\boldsymbol{\theta}_p = (\boldsymbol{\omega}_{\bullet p}, \omega_{pp})$ and $z =$ collection of all other latent variables.

- Wang (2012, BA) showed in the context of sampling that $f(\boldsymbol{\theta}_p \mid \mathbf{y}, z) = f(\boldsymbol{\omega}_{\bullet p}, \omega_{pp} \mid \mathbf{y}, z) = f(\boldsymbol{\omega}_{\bullet p} \mid \mathbf{y}, z) f(\omega_{pp} \mid \boldsymbol{\omega}_{\bullet p}, \mathbf{y}, z)$ decomposes as (normal $\times$ gamma) under suitable priors on $\boldsymbol{\Omega}_{p\times p}$.

- 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

$$\log f(\mathbf{y}_{1:p}) \quad = \quad \log f(\mathbf{y}_{1:p} \mid \boldsymbol{\theta}_p) + \log f(\boldsymbol{\theta}_p) - \log f(\boldsymbol{\theta}_p \mid \mathbf{y}_{1:p}).$$

- Slightly rewrite:

$$\begin{aligned}\log f(\mathbf{y}_{1:p}) &= \log f(\mathbf{y}_p \mid \mathbf{y}_{1:p-1}, \boldsymbol{\theta}_p) + \log f(\mathbf{y}_{1:p-1} \mid \boldsymbol{\theta}_p) + \log f(\boldsymbol{\theta}_p) - \log f(\boldsymbol{\theta}_p \mid \mathbf{y}_{1:p}) \\ &:= \mathrm{I}_p + \mathrm{II}_p + \mathrm{III}_p - \mathrm{IV}_p.\end{aligned}$$

- We can evaluate the partial likelihood I_p using

$$\mathbf{y}_p \mid \mathbf{y}_{1:p-1}, \boldsymbol{\theta}_p \sim \mathcal{N}(-\mathbf{y}_{1:p-1}\boldsymbol{\omega}_{\bullet p} / \omega_{pp},\ 1/\omega_{pp}),$$

- 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 II is telescoping. We have:

$$\begin{aligned}\mathrm{II}_p &= \log f(\mathbf{y}_{1:p-1} \mid \boldsymbol{\theta}_p) \\ &= \log f(\mathbf{y}_{p-1} \mid \mathbf{y}_{1:p-2}, \boldsymbol{\theta}_p, \boldsymbol{\theta}_{p-1}) + \log f(\mathbf{y}_{1:p-2} \mid \boldsymbol{\theta}_p, \boldsymbol{\theta}_{p-1}) \\ &\quad + \log f(\boldsymbol{\theta}_{p-1} \mid \boldsymbol{\theta}_p) - \log f(\boldsymbol{\theta}_{p-1} \mid \mathbf{y}_{1:p-1}, \boldsymbol{\theta}_p) \\ &:= \mathrm{I}_{p-1} + \mathrm{II}_{p-1} + \mathrm{III}_{p-1} - \mathrm{IV}_{p-1}.\end{aligned}$$

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

$$\widetilde{\boldsymbol{\Omega}}_{(p-1)\times(p-1)} = \boldsymbol{\Omega}_{(p-1)\times(p-1)} - \frac{\boldsymbol{\omega}_{\bullet p}\boldsymbol{\omega}_{\bullet p}^T}{\omega_{pp}} := \begin{bmatrix} \widetilde{\boldsymbol{\Omega}}_{(p-2)\times(p-2)} & \widetilde{\boldsymbol{\omega}}_{\bullet(p-1)} \\ \widetilde{\boldsymbol{\omega}}_{\bullet(p-1)}^T & \widetilde{\omega}_{(p-1)(p-1)} \end{bmatrix}.$$

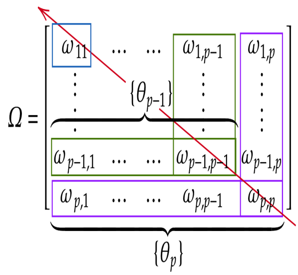
  Then $\widetilde{\boldsymbol{\Omega}}_{(p-1)\times(p-1)}$ is p.d. and $(\mathbf{y}_{1:p-1} \mid \boldsymbol{\theta}_p,\, \boldsymbol{\Omega}_{(p-1)\times(p-1)}) \sim \mathcal{N}(0, \widetilde{\boldsymbol{\Omega}}^{-1}_{(p-1)\times(p-1)})$.

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

$$\mathbf{y}_{p-1} \mid \mathbf{y}_{1:p-2},\, \boldsymbol{\theta}_p,\, \boldsymbol{\theta}_{p-1} \sim \mathcal{N}(-\mathbf{y}_{1:p-2}\widetilde{\boldsymbol{\omega}}_{\bullet(p-1)}\,/\widetilde{\omega}_{(p-1)(p-1)},\ 1/\widetilde{\omega}_{(p-1)(p-1)}).$$

Overall strategy

(a)



(b)

$$
\begin{array}{ccccccccc}
\log f(\mathbf{y}_{1:p}) & = & \mathrm{I}_p & + & \cancel{\mathrm{II}_p} & + & \mathrm{III}_p & - & \mathrm{IV}_p \\
\cancel{\mathrm{II}_p} & = & \mathrm{I}_{p-1} & + & \cancel{\mathrm{II}_{p-1}} & + & \mathrm{III}_{p-1} & - & \mathrm{IV}_{p-1} \\
\cancel{\mathrm{II}_{p-1}} & = & \mathrm{I}_{p-2} & + & \cancel{\mathrm{II}_{p-2}} & + & \mathrm{III}_{p-2} & - & \mathrm{IV}_{p-2} \\
\vdots & = & \vdots & & \vdots & & \vdots & & \vdots \\
\vdots & = & \vdots & + & \vdots & + & \vdots & - & \vdots \\
\vdots & = & \vdots & & \vdots & & \vdots & & \vdots \\
\cancel{\mathrm{II}_3} & = & \mathrm{I}_2 & + & \cancel{\mathrm{II}_2} & + & \mathrm{III}_2 & - & \mathrm{IV}_2 \\
\cancel{\mathrm{II}_2} & = & \mathrm{I}_1 & + & (\mathrm{II}_1 = 0) & + & \mathrm{III}_1 & - & \mathrm{IV}_1 \\
\hline
\log f(\mathbf{y}_{1:p}) & = & \sum_{j=1}^{p} \mathrm{I}_j & + & 0 & + & \sum_{j=1}^{p} \mathrm{III}_j & - & \sum_{j=1}^{p} \mathrm{IV}_j \\
\hline
\end{array}
$$

Figure: (a) Decomposition of $\boldsymbol{\Omega}_{p\times p}$. Purple, green and blue blocks denote $\boldsymbol{\theta}_p, \boldsymbol{\theta}_{p-1}$ and finally $\boldsymbol{\theta}_1 = \omega_{11}$. Red arrow denotes how the algorithm proceeds, fixing one row/column at a time, and (b) the telescoping sum giving the log-marginal $\log f(\mathbf{y}_{1:p})$.

- Run Chib p times, adjusting $\boldsymbol{\Omega}$ each time. In each equation, evaluate only I, III and IV. Eliminate II via telescoping sum.

A demonstration on Wishart (known marginal)

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

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

- The closed form expression for the marginal provides an *oracle*.

Computing III_p $(= \log f(\boldsymbol{\theta}_p))$

- Recall, $\boldsymbol{\theta}_p = (\boldsymbol{\omega}_{\bullet p}\,, \omega_{pp})$.

- If $\boldsymbol{\Omega} \sim \mathcal{W}_p(\mathbf{I}_p, \alpha)$ then $f(\boldsymbol{\omega}_{\bullet p}\,, \omega_{pp}) = f(\boldsymbol{\omega}_{\bullet p} \mid \omega_{pp}) f(\omega_{pp})$, where,

  $\boldsymbol{\omega}_{\bullet p} \mid \omega_{pp} \sim \mathcal{N}(0, \omega_{pp}\mathbf{I}_{p-1})$, $\omega_{pp} \sim \mathrm{Gamma}(\text{shape} = \alpha/2, \text{rate} = 1/2)$.

- Computing III_p is easy: normal $\times$ gamma.

Computing $\mathrm{IV}_p \, (= \log f(\boldsymbol{\theta}_p \mid \mathbf{y}_{1:p}))$

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

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{(p-1)\times(p-1)} & \mathbf{s}_{\bullet p} \\ \mathbf{s}_{\bullet p}^T & s_{pp} \end{bmatrix}, \, \boldsymbol{\beta}_{\bullet p} = \boldsymbol{\omega}_{\bullet p}\,, \, \gamma_{pp} = \omega_{pp} - \boldsymbol{\omega}_{\bullet p}^T \boldsymbol{\Omega}_{(p-1)\times(p-1)}^{-1} \boldsymbol{\omega}_{\bullet p}.$$

- Key result of Wang (2012, BA):

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

where $\mathbf{C} = \{(s_{pp}+1)\boldsymbol{\Omega}_{(p-1)\times(p-1)}^{-1}\}^{-1}$

- Allows computation of IV_p using Chib's two block strategy.

Computing $\mathrm{III}_{p-1}, \ldots, \mathrm{III}_1$ and $\mathrm{IV}_{p-1}, \ldots, \mathrm{IV}_1$

- Same strategy as in going from I_p to I_{p-1}.
- Proceed backwards starting from the pth row. At each step, adjust the upper left sub-matrix $\boldsymbol{\Omega}_{j\times j}$ via IPS:

for (j=p-1,. . . , 1) **do**

Update $\boldsymbol{\Omega}_{j\times j} \leftarrow \boldsymbol{\Omega}_{j\times j} - \frac{\boldsymbol{\omega}_{\bullet(j+1)}\boldsymbol{\omega}^T_{\bullet(j+1)}}{\omega_{j+1,j+1}}$.

end for

- Calculate III_j and IV_j with the updated $\boldsymbol{\Omega}_{j\times j}$.

Results for Wishart

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

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:

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

- Two examples (global-local shrinkage priors):
  - Bayesian graphical lasso (BGL):

$$f(\omega_{ij} \mid \lambda) = (\lambda/2) \exp(-\lambda|\omega_{ij}|)$$
$$\Updownarrow \text{ (Andrews and Mallows, 1974)}$$
$$\omega_{ij} \mid \tau_{ij}, \lambda \sim \mathcal{N}(0, \tau_{ij}), \quad \tau_{ij} \mid \lambda \sim \mathrm{Exp}(\lambda^2/2).$$

  - 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 ω_{ij} terms are normal conditional on τ_{ij}.
- Similarly, the diagonal ω_{jj} are exponential.
- The presence of these mixing τ_{ij} variables is the ONLY difference with the Wishart case for our purposes.
- The τ_{ij} terms can be sampled easily.
- MAIN IDEA: Absorb the τ_{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 $\mathrm{IV}_p \, (= \log f(\boldsymbol{\theta}_p \mid \mathbf{y}_{1:p}))$

- We have

$$f(\boldsymbol{\beta}_{\bullet p}, \gamma_{pp} \mid \boldsymbol{\tau}_{\bullet p}, \boldsymbol{\Omega}_{(p-1)\times(p-1)}, \mathbf{y}_{1:p}) = \mathcal{N}(\boldsymbol{\beta}_{\bullet p} \mid -\mathbf{C}\mathbf{s}_{\bullet p}, \mathbf{C}) \times \mathrm{Gamma}\left(\gamma_{pp} \mid \frac{n}{2} + 1, \frac{s_{pp} + \lambda}{2}\right),$$

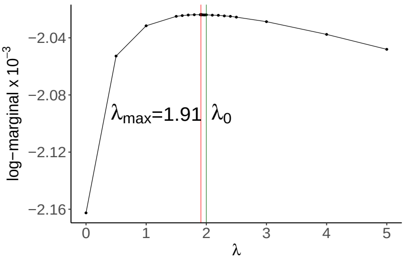
where $\mathbf{C} = \{\mathrm{diag}^{-1}(\boldsymbol{\tau}_{\bullet p}) + (s_{pp} + \lambda)\boldsymbol{\Omega}^{-1}_{(p-1)\times(p-1)}\}^{-1}$

- Recall, for Wishart we had

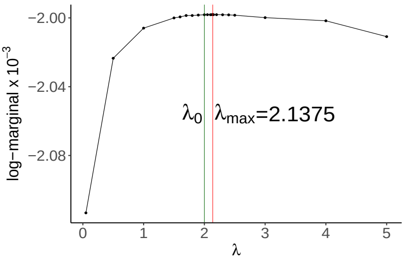
$$f(\boldsymbol{\beta}_{\bullet p}, \gamma_{pp} \mid \boldsymbol{\Omega}_{(p-1)\times(p-1)}, \mathbf{y}_{1:p}) = \mathcal{N}(\boldsymbol{\beta}_{\bullet p} \mid -\mathbf{C}\mathbf{s}_{\bullet p}, \mathbf{C}) \times \mathrm{G}\left(\gamma_{pp} \mid \frac{n + \alpha - p - 1}{2} + 1, \frac{s_{pp} + 1}{2}\right).$$

where $\mathbf{C} = \{(s_{pp} + 1)\boldsymbol{\Omega}^{-1}_{(p-1)\times(p-1)}\}^{-1}$.

Results



(a) BGL



(b) GHS

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

	λ	0.05	1	$2\ (= \lambda_0)$	3	4	5
$\log \mathrm{BF}$	BGL	138.84	7.86	0.18	4.98	13.9	24.34
	GHS	115.31	7.89	0.12	1.79	3.63	12.83

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 $\boldsymbol{\Omega}$ are scale mixtures of normal and (b) the diagonals of $\boldsymbol{\Omega}$ 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 $\mathbf{y} \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1})$
- This is equivalent to $\mathbf{y} \mid \tau \sim \mathcal{N}(\boldsymbol{\mu}, \tau^{-1}\boldsymbol{\Omega}^{-1}),\ \tau \sim \mathrm{Gamma}(\nu/2, \nu/2)$.
- Should be possible to absorb the τ in the likelihood into Chib's z and proceed.

Main references

- **Bhadra, A.**, Sagar, K., Banerjee, S. and Datta, J. (2022+). Graphical Evidence. (submitted). [arXiv:2205.01016]
  - Code: `https://github.com/sagarknk/Graphical_Evidence`
- Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313–1321.
- Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. Bayesian Analysis **7**, 867–886.