Bayes Model Selection with Path Sampling: Factor Models

Ritabrata Dutta and Jayanta K Ghosh
Aalto University, Finland and Purdue University, USA

31/12/12

Factor Models in Applications

Factor Models and Factor analysis

- originated in psychometrics, and is used in behavioral sciences, social sciences and other applied sciences etc. that deal with large quantities of data.
- recently been used in social science applications, genomic applications (West, 2003 ; Carvalho et al., 2008) and epidemiologic studies (Sanchez et al, 2005)
- is popular for its ability to model the sparse covariance structures in the high-dimensional problems these days.

Factor Models

- A factor model, with k factors is defined as n i.i.d. observed p-dimensional r.v.'s

$$y_i = \Lambda \eta_i + \epsilon_i, \epsilon_i \sim N_p(0, \Sigma),$$

where Λ is a $p \times k$ matrix of factor loadings,

$$\eta_i = (\eta_{i1}, \ldots, \eta_{ik})' \sim N_k(0, I_k)$$

is a vector of standard normal latent factors, and ϵ_i is the residual with diagonal covariance matrix
$\Sigma = diag(\sigma_1^2, \ldots, \sigma_p^2)$.

Factor Models

- Thus we may write the marginal distribution of y_i as $N_p(0, \Omega)$, $\Omega = \Lambda\Lambda' + \Sigma$.
- This model implies that the sharing of common latent factors explains the dependence in the outcomes.

Factor Models

- To solve **non-identifiability under orthogonal rotation**, it is customary to assume that Λ has a full-rank lower triangular structure, restricting the number of free parameters in Λ and Σ to $q = p(k+1) - k(k-1)/2$, where k must be chosen so that $q \leq p(p+1)/2$.
- The reciprocal of diagonal entries of Σ form the **precision vector** here.

Factor Models : Model Selection ?

- For a given data set $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, which of the following models are true ?
- Probable Models : $M_1, M_2, \ldots, M_m$ when M_k is a Factor Model with k-factors.
- Bayes Answer : The model M_k with highest posterior probability $P(M_k|\mathbf{y})$ given the data $\mathbf{y}$ will be our chosen model, while we assume equal prior probability for each of the m-models.
- So we need to compute $P(M_k|\mathbf{y}),\ \forall k$

Factor Models : Specification of Priors

- Prior elicitation is not common.
- commonly used in practice due to conjugacy :
  1. Truncated normal priors for the diagonal elements of Λ,
  2. normal priors for the lower triangular elements,
  3. and inverse-gamma priors for $\sigma_1^2, \ldots, \sigma_p^2$,

Factor Models : Specification of Priors

- **Diffuse Prior** : Gelman (2006) introduced a new class of **diffuse** priors for the factor loadings that have good mixing properties through parameter expansion.
- Using **parameter expansion** Ghosh and Dunson (2009) induced **t or folded-t** priors depending on sign constraints on the loadings.
- We use the same family (t or folded-t) of priors but consider a whole range of **many degrees of freedom** going all the way to the normal and use the similar Gibbs sampler as in Ghosh and Dunson (2009).

Bayes Model Selection and Bayes Factor

- **Marginal likelihood** under model M_k,

$$m_k(y) = P(y|M_k)$$

is obtained by integrating the likelihood

$$\Pi_i N_p(y_i; 0, \Lambda^{(k)}\Lambda^{(k)'} + \Sigma)$$

across the prior distribution for the factor loadings $\Lambda^{(k)}$ and residual variance Σ.

- Notice the marginal likelihood $m_k(y)$ is the **normalizing constant** of the posterior distribution under model M_k.

Bayes Model Selection and Bayes Factor

- **Bayes Factor** $BF_{k:j}$ is defined as

$$BF_{k:j} = m_k(y)/m_j(y)$$

or the **ratio of the normalizing constants** of the posterior distributions under model M_k and M_j.

Bayes Model Selection and Bayes Factor

- Under the assumption of equal prior probability for the models, posterior probability for any model M_k can be written as

$$P(M_k|y) = \frac{BF_{k:j}}{\sum_{l=1}^{m} BF_{l:j}}.$$

- So computation of $P(M_k|\mathbf{y})$ is equivalent to the computation of **Bayes Factor**.

Bayes Model Selection and Bayes Factor

- For Factor Models estimation of Bayes Factors between two consecutive models

$$BF_{h;h-1} = \frac{m_h(x)}{m_{h-1}(x)}$$

  where $m_h(x)$ is the marginal under the model having h latent factors, is sufficient.
- So we can see estimation of Bayes Factor as estimation of ratio of normalizing constants under two densities. (Z_1/Z_0 when Z_i is the normalizing constant of the posterior distribution under model M_i.)

Different Methods in Estimation of Bayes Factor

- **Importance Sampling** : Easy to implement and shows moderately good results in choosing the correct model but do have a very large MCMC-variance.
- **Newton-Raftery approximation (BICM)** and **Laplace/BIC type approximation (BICIM)** : Shows significantly less amount of MCMC-variance but not very reliable in choosing the correct model.
- **RJMCMC** : Difficult to implement and lots of convergence issues regarding MCMC.
- **Path-Sampling** : Along the line of Importance Sampling and shows good results in choosing the correct model.

Path Sampling (Gelman and Meng (1998))

- Construct a path $f_t : t \in [0,1]$ connecting f_0 and f_1, the non-normalized density of M_0 and M_1.
- Suppose the path is given by $p_t : t \in [0,1]$ where for each t, p_t is a probability density, having the following identity.

$$p_t(\theta) = \frac{1}{z_t} f_t(\theta), \tag{1}$$

where f_t is an unnormalized density and $z_t = \int f_t(\theta) d\theta$ is the normalizing constant.

PathSampling

- Taking the derivative of the logarithm on both sides, we obtain the following identity under the assumption of interchangeability of the order of integration and differentiation :

$$\begin{aligned} \frac{d}{dt} log(z_t) &= \int \frac{1}{z_t} \frac{d}{dt} f_t(\theta) \mu(d\theta) \\ &= E_t[\frac{d}{dt} \log f_t(\theta)] \\ &= E_t[U(\theta, t)] \end{aligned} \tag{2}$$

where the expectation E_t is taken with respect to $p_t(\theta)$ and $U(\theta, t) = \frac{d}{dt} \log f_t(\theta)$.

PathSampling

- Now integrating (2) from 0 to 1 gives the log of the ratio of the normalizing constants, i.e. log BF in the context of model selection :

$$log[\frac{Z_1}{Z_0}] = \int_0^1 E_t[U(\theta, t)]dt \tag{3}$$

PathSampling : Computation

- To approximate the integral we discretize the path with k points

$$t_{(0)} = 0 < t_{(1)} < \ldots < t_{(k)} = 1$$

and draw m MCMC samples converging to $p_t(\theta)$ at each of these k points.

- Then estimate $E_t[U(\theta, t)]$ by $\frac{1}{m} \sum U(\theta^{(i)}, t)$ where $\theta^{(i)}$ is the MCMC output.

PathSampling : Computation

- To estimate the final log Bayes factor, commonly numerical integration schemes are used.
- **Parallelization :** It is clear that MCMC at different points "t" on the path can be done in parallel. We have used this both for PS and for our modification of it.

Path Sampling : Different Paths

- **Arithmetic Mean Path (AMP)** : defined by the mean $f_t = tf_0 + (1-t)f_1$ of the densities of two competing models for each model $M_t : t \in (0,1)$ along the path.
- **Geometric Mean Path (GMP)** : defined by the mean $f_t = f_0^t f_1^{(1-t)}$ of the densities of two competing models for each model $M_t : t \in (0,1)$ along the path.
- **Parametric Arithmetic Mean Path (PAMP)** : One more common path is obtained by assuming a specific functional form f_θ for the density and then constructing the path in the parametric space $\theta \in \Theta$ of the assumed density. If $\theta_t = t\theta_0 + (1-t)\theta_1$, then f_{t,θ_t} is the density of the model Mt, where $f_{0,\theta_0} = f_0$ and $f_{1,\theta_1} = f_1$.

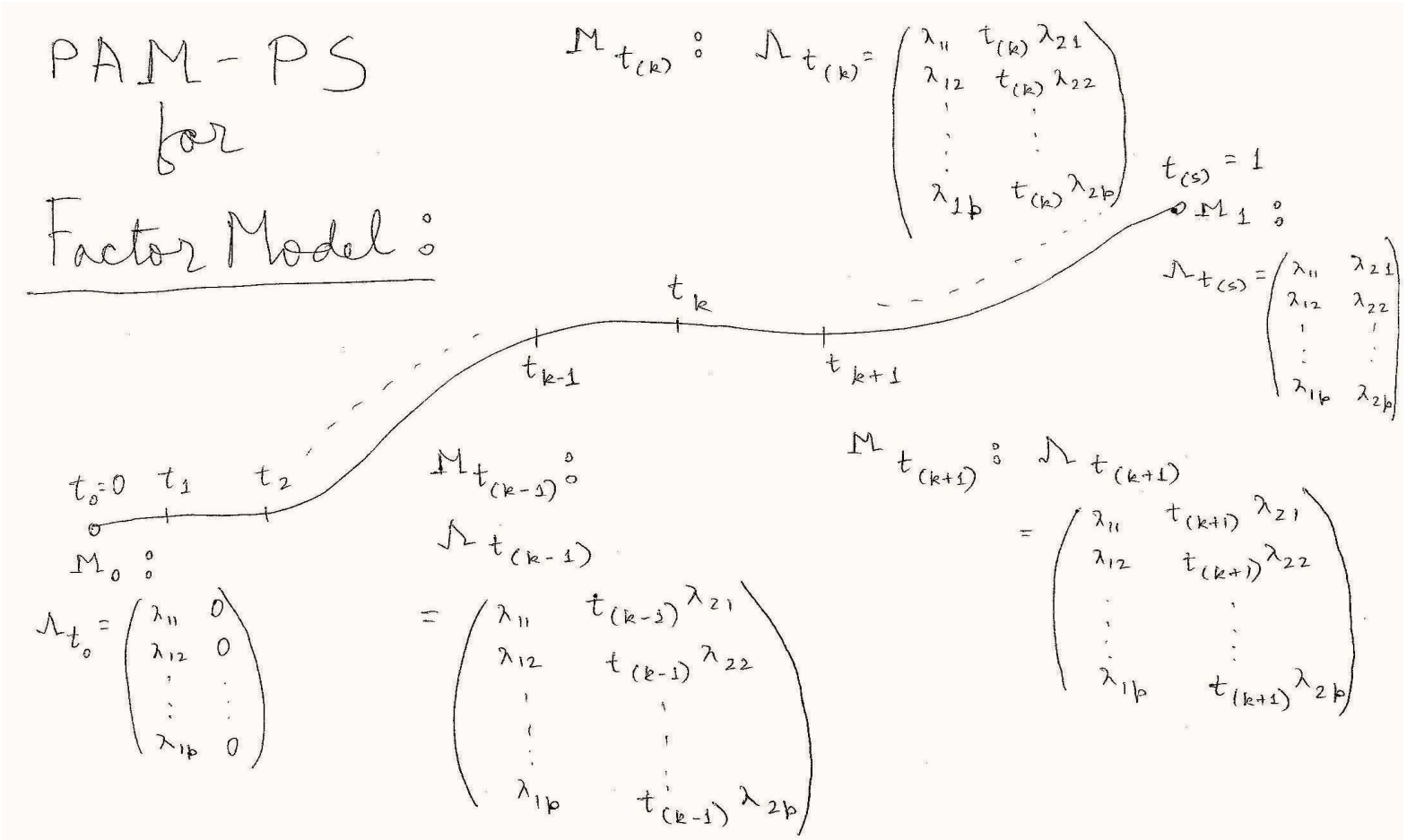
Path Sampling for Factor Models : PAM Path

- Here we define M_0 and M_1 to be the two models corresponding to the factor model with factors $h-1$ and h, respectively and then connect them by the path

$$M_t : y_i = \Lambda_t \eta_i + \epsilon_i, \Lambda_t = (\lambda_1, \lambda_2, \ldots, \lambda_{h-1}, t\lambda_h)$$

  where λ_j is the j-th column of the loading matrix. So for t=0 and t=1 we get the models M_0 and M_1.
- The likelihood function at grid point t is a MVN which is denoted as $f(Y|\Lambda, \Sigma, \eta, t)$.

PAM Path for Factor Models



Path Sampling for Factor Models : Estimation of log BF

- Under independent priors, our score function is :

$$U(\Lambda, \Sigma, \eta, Y, t) = \sum_{i=1}^{n} (y_i - \Lambda_t \eta_i)' \Sigma^{-1} (0^{p \times (h-1)}, \lambda_h) \eta_i. \quad (4)$$

Path Sampling for Factor Models : Estimation of log BF

- For fixed and ordered grid points along the path $t_{(0)} = 0 < t_{(1)} < \ldots < t_{(S)} < t_{(S+1)} = 1$: We simulate m samples of $(\Lambda_{t_s,i}, \Sigma_i, \eta_i : i = 1, \ldots, m)$ from the posterior distribution of $(\Lambda_{t_s}, \Sigma, \eta)$ at the point $0 \leq t_s \leq 1$ and use them to estimate
  $\widehat{E}_s(U) = \frac{1}{m} \sum U(\Lambda_{t_s,i}, \Sigma_i, \eta_i, y), \ \ \forall s = 1, \ldots, S+1.$

Path Sampling for Factor Models : Estimation of log BF

- Our path sampling estimate for the log Bayes factor is

$$log(\widehat{BF}_{h:h-1}) = \frac{1}{2}\sum_{s=0}^{S}(t_{s+1} - t_s)(\widehat{E}_{s+1}(U) + \widehat{E}_s(U)). \qquad (5)$$

Path Sampling for Factor Models : An Easy Theorem

- The theorem below verifies the regularity conditions of Path Sampling for Factor Models. For PS to succeed we also need convergence of MCMC at each point in the path. That will be taken up after the theorem.
- **Theorem** : Consider Path Sampling for factor models with parametric arithmetic mean path (PAMP), likelihood as given above for factor models and **a proper prior for which the score function is integrable w.r.t. the prior**,
  1. The interchangeability of integration and differentiation in (2) is valid.
  2. $E_t(U)$ is finite as $t \to 0$.
  3. The path sampling integral for factor models, namely (3), is finite.

Path Sampling for Factor Models : Heuristic

- We will look into the **spread and mixing of the MCMC sample** first
- **Difficult to visualize** : As the parameters are **high-dimensional**
- So look into the **log-likelihood values** of those parameters in search of some patterns

Path Sampling for Factor Models : Heuristic

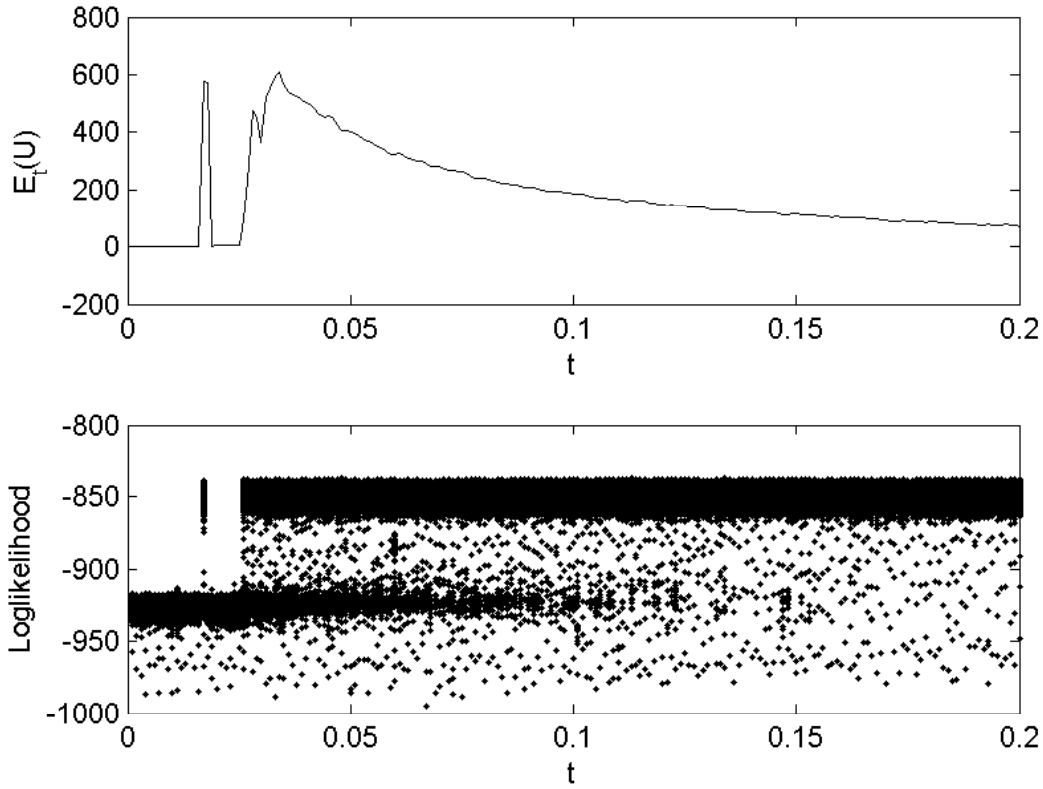


FIGURE : $E_t(U)$ and *Loglikelihood* for prior t_{10} in the range $t \in [0, .2]$, 2-factor model is true.

Path Sampling for Factor Models : Heuristic

- We see two clusters in the figure.
- One around the **points maximizing likelihood** under model M_1.
- Another around **prior mode and the mle under** M_0, namely $\lambda_h = 0$.

Path Sampling for Factor Models : Heuristic

- We expect to see **a range (say** $[t_1, t_2]$**)** near zero showing a **conflict between prior and maximized likelihood**. Definitely the points t_1 and t_2 are not well-specified, but we treat them as such so as to understand some underlying issues of mixing and convergence here showing a **lot of fluctuations from MCMC to MCMC.**
- On the set of points $t > t_2$ the MCMC samples are expected to be **around the points maximizing likelihood**
- Whereas for $t < t_1$ they will be nearly **zero** due to the concentration around a value λ_h which is both **prior mode and the mle under** M_0, namely $\lambda_h = 0$.

Path Sampling for Factor Models : Heuristic

The **width of the zone of conflict** (here $t_2 - t_1$) will shrink, if we have a relatively strong decaying tail of the prior.

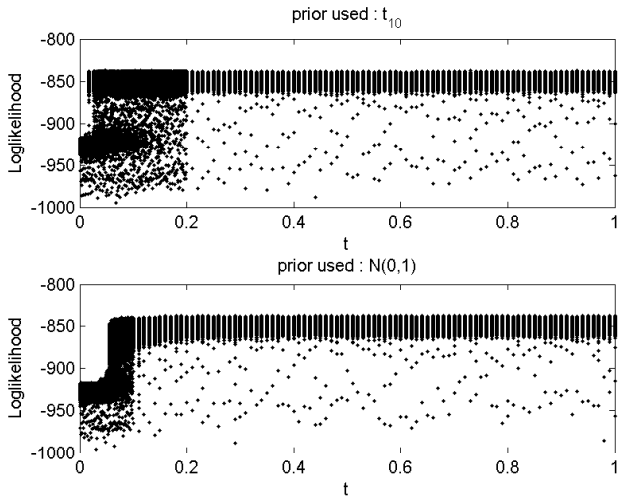


FIGURE : $E_t(U)$ and *Loglikelihood* for prior t_{10} in the range $t \in [0, .2]$, 2-factor model is true.

Path Sampling for Factor Models : Heuristic

- If the prior fails to have the required finite moment in the theorem, the posterior will also be likely to have large values for moments, which may cause convergence problems for the MCMC. That's why we chose a prior making the score function integrable.
- In the proof of the Thoerem, we have assumed the first two moments of the prior to be finite. **In most numerical work our prior is a t with 5 to 10 d.f.**

Path Sampling for Factor Models : Heuristic

- When the complex model is true, the **loglikelihood between the two models differ by an order of magnitude.**
- For heavy-tailed priors we may see these above mentioned fluctuations for a longer range, causing a **loss of mass from the final integration**.
- These problems are aggravated by the **high-dimension of the problem and the diffuse spread of the prior on the high-dimensional space.**
- This may mean the usual BF estimated by PS will be off by an order of magnitude.

Basic Facts for Computation

- We use a 2-factor model and 1-factor model as our complex model M_1 and simpler model M_0 correspondingly to demonstrate the underlying issues.
- The loading parameters and the diagonal entries of the Σ matrix are given in following Tables.

Λ	Factor 1	.89	0	.25	0	.8	0	.5
	Factor 2	0	.9	.25	.4	0	.5	0
Σ		.2079	.19	.15	.2	.36	.1875	.1875

TABLE : Loading Factors and Diagonal Entries of Σ used for simulation

Basic Facts for Computation

- Error in estimation of the BF or the discrepancy between different methods tends to be relatively large, if one of the following is true :
  1. the data has come from the complex model rather than the simpler model
  2. the prior is relatively diffuse
  3. the value of the precision parameters are relatively small

Issues in Complex (2-factor) Model

PS using grid size .01				
t_1	t_5	t_{10}	t_{90}	normal
2.62	14.42	22.45	70.20	70.25
3.67	11.90	21.39	68.70	68.72
3.00	13.43	21.31	47.06	47.21
4.29	13.17	18.49	48.03	48.13
4.20	13.11	18.48	47.70	47.74

TABLE : PAM-PS : Dependance of $\log BF_{21}$ over prior, 2-factor model true.

Issues in Complex (2-factor) Model

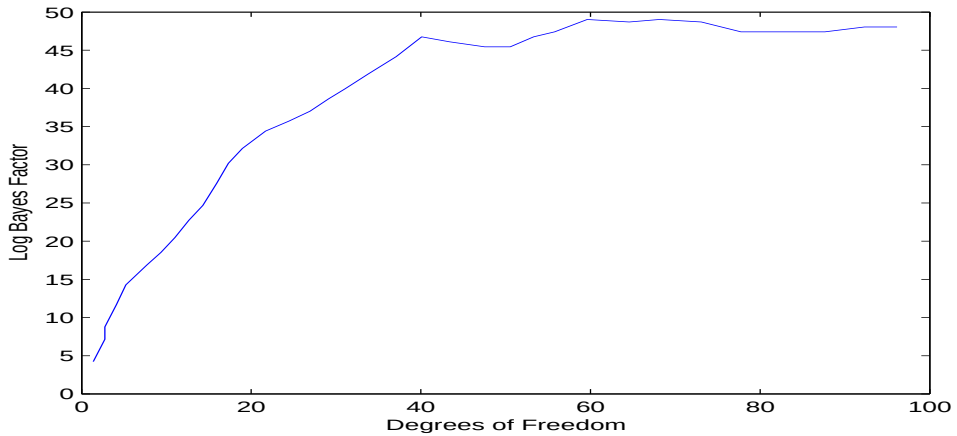


FIGURE : Dependance of $\log BF_{21}$ over prior for 3rd Data set.

Issues in Complex (2-factor) Model

Grid Size		.01	.001
MCMC-size	Prior	Data 2	Data 2
5000	t_{10}	21.26 (1.39)	21.2625 (1.29)
	N(0,1)	66.89 (4.16)	67.2158 (3.18)
50000	t_{10}	23.7189 (1.8)	23.5783 (.52)
	N(0,1)	68.1035 (3.72)	68.23 (3.66)

TABLE : PAM-PS : Dependence of log BF_{21} (MCMC-variance) estimates over MCMC size, 2-factor model true

Issues in Complex (2-factor) Model

- We can see the estimate of the BF changing with the change in the pattern of the tail of the prior.
- Major increase of MCMC-size and finer grid-size reduces the MCMC-variance of the estimator. The difference between the mean values of BF estimated by t_{10} and N(0,1) differ by an order of magnitude.

Issues in Simpler (1-factor) Model

Grid Size		.01	.001
MCMC-size	Prior	Data 1	Data 1
5000	t_{10}	-4.26 (.054)	-4.2702 (.044)
	N(0,1)	-4.62 (.052)	-4.6045 (.051)
50000	t_{10}	-4.2457 (.012)	-4.246 (.007)
	N(0,1)	-4.6064 (.0065)	-4.6267 (.005)

TABLE : PAM-PS : Dependence of log BF_{21} (MCMC-variance) estimates over MCMC size, while 1-factor model true

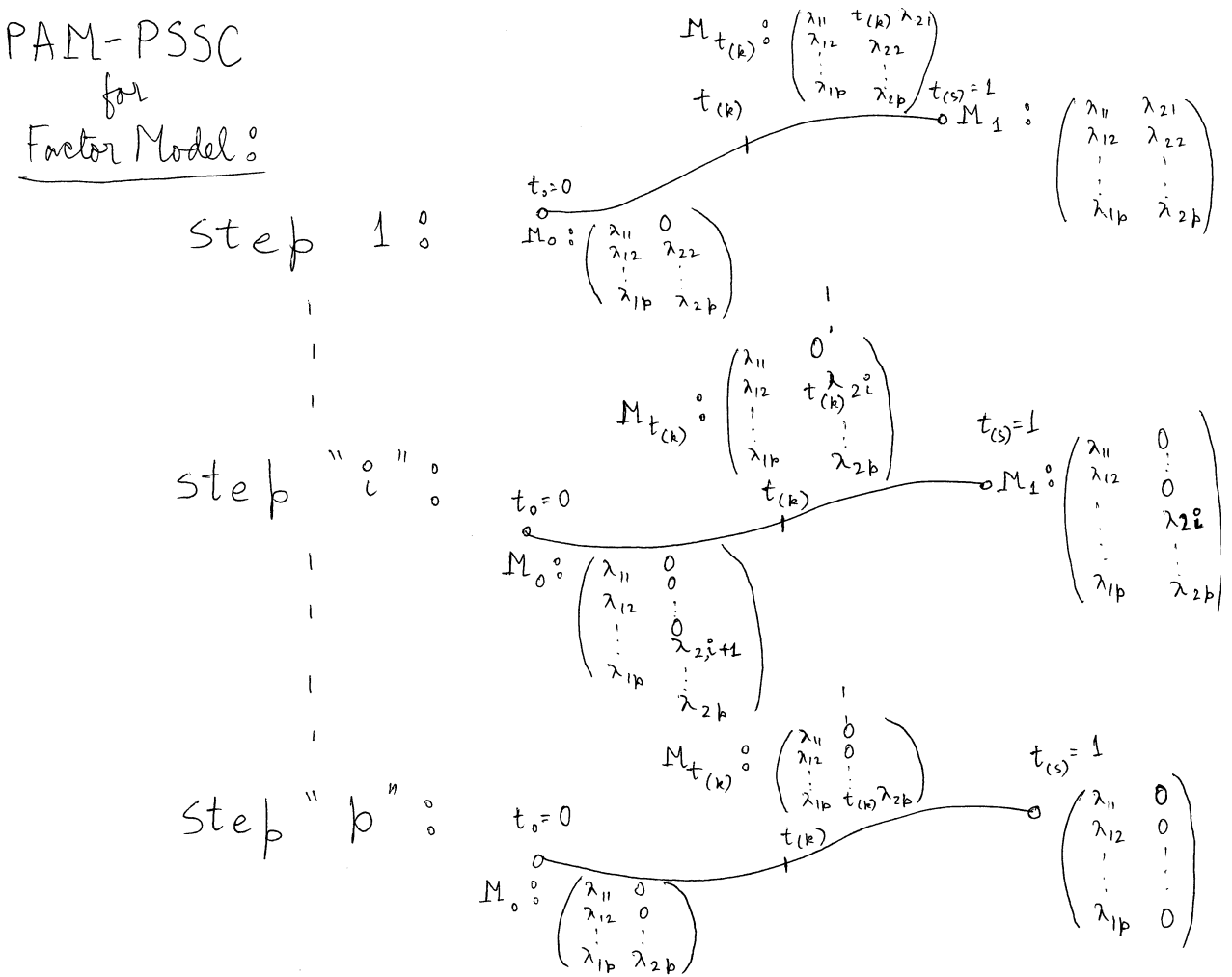
Issues in Simpler (1-factor) Model

- This table shows us that the MCMC-variance improves with the finer grid-size and large MCMC-size as expected, but the estimated values of BF_{21} remain mostly same.
- As noted earlier, PS chooses the correct model 100% of the time when M_0 is true.

Motivation for this new solution (PSSC)

- So we see when **two models are close in some sense**, we expect their likelihood ratio will not fluctuate widely.
- To solve this problem without having to give up our **diffuse prior**, we try to reduce the problem to a **series of one-dimensional problems so that the competeting models are close to each other**.
- We calculate the Bayes Factor by using the pathsampling step for every single parameter that may be zero, keeping others fixed.

Paths used for PSSC



How does PSSC work ?

- More formally, if we consider λ_2 as a p-dimensional vector, then M_0 and M_1 differ only in the last $p-1$ parameters, as λ_{21} is always zero due to upper-triangular condition.
- We consider p models $M_i' : i = 1, \ldots, p$, where for model M_i' we have first i parameters of λ_2 being zero correspondingly.

How does PSSC work ?

- If we define $BF'_{i,i+1} = \frac{m_i(x)}{m_{i+1}(x)}$, when $m_i(x)$ is the marginal for the model M'_i then,

$$\log BF_{21} = \sum_{i=1}^{p-1} \log BF'_{i,i+1}.$$

- So we perform $p-1$ pathsampling computations to estimate $\log BF'_{i,i+1}, \forall i = 1, \ldots, p-1$.

Comparison between PS and PSSC

True Model	MCMC Size	PS-SC (t_{10})	PS (t_{10})	PS (N(0,1))
1-factor	5000	-8.09 (.013)	-4.26 (.054)	-4.62 (.052)
1-factor	50000	-8.0892 (.0067)	-4.24 (.012)	-4.60 (.0065)
2-factor	5000	80.14 (.67)	21.26 (1.39)	66.89 (3.8)
2-factor	50000	80.75 (.83)	23.7189 (1.8)	68.10 (3.88)

TABLE : log BF_{21} (variance over MCMC sample) estimated by PAM-PS-SC

Comparison between PS and PSSC

Model	True Model	Data	PS-SC (t_{10})	PS (t_{10})
Model 1	1-factor	Data 1	-8.09 (.012)	-3.84 (.055)
Precision $\in [2.77, 6.55]$	2-factor	Data 2	71.59 (.66)	19.81 (1.38)
Model 2	1-factor	Data 1	-11.01 (.0066)	-3.09 (.0277)
Precision $\in [1.79, 2.44]$	2-factor	Data 2	51.41 (.3658)	2.8 (1.9104)
Model 3	1-factor	Data 1	-5.13 (.0153)	-2.6 (.0419)
Precision $\in [1.36, 1.66]$	2-factor	Data 2	3.975 (.0130)	2.2 (.3588)

TABLE : $\log BF_{21}$ (MCMC-variance) estimation by PS-SC : Effect of Precision Parameter

Comparison between PS and PSSC (Poor Mixing in PS)

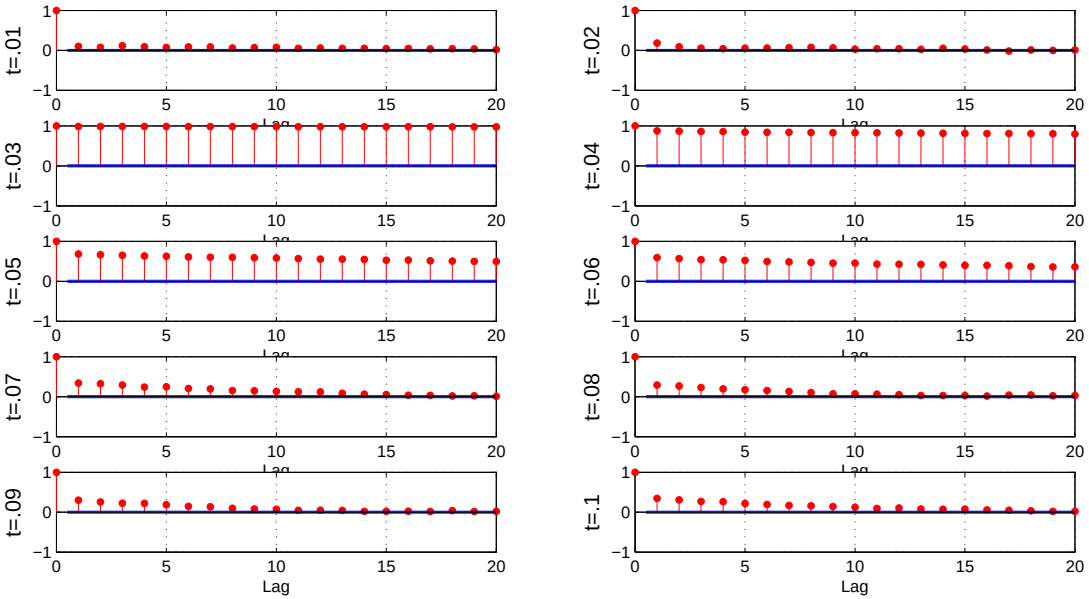


FIGURE : Autocorrelation for λ'_{22} for different values of t near t=0 (MCMC size used 50,000), using PS.

Comparison between PS and PSSC (Better Mixing in PSSC)

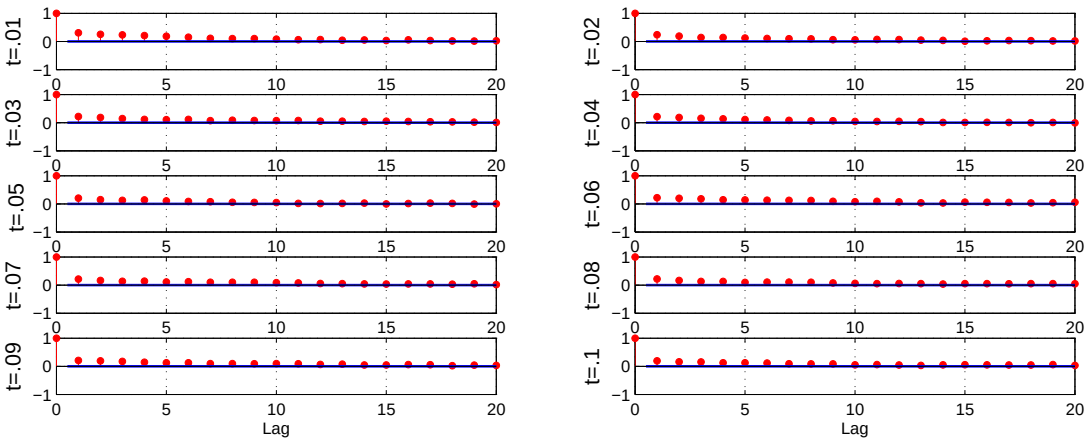


FIGURE : Autocorrelation for λ'_{22} for different values of t near t=0 (MCMC size used 50,000), using PS-SC.

Applications on some Real Data

Bayes factor	PS-SC	BICM	BICIM
$logBF_{21}$	4.8	26.34	21.57
$logBF_{32}$	10.52	-3.14	-10.01
$logBF_{43}$	-3.28		

TABLE : Rodant Organ Weight Data (p=6, n=60) : Comparison of log Bayes factor

Bayes factor	PS-SC	BICM	BICIM
$logBF_{21}$	122.82	205.27	188.19
$logBF_{32}$	35.27	71.05	35.5
$logBF_{43}$	-10.7	23.16	7.55
$logBF_{54}$	-33.32	-4.63	-25.51
$logBF_{65}$	-16.7	-17.32	-43.21

TABLE : 26-variable Psychological data (p=26, n=300) : Comparison of log Bayes factor

Conclusion

- The models chosen by PS-SC and the other methods are close, but as expected differ a lot in their estimate of BFs.
- The two heuristic Laplace approximationsseem to be good as a preliminary searching method to narrow the field of plausible models before using PS-SC.
- We have studied PS for Factor Models and noticed **"The problem is worse the more the two models differ as when a very high dimensional model is being compared to a low dimensional model"**.
- It is our belief that the above insights as to when things will tend to go wrong and when not, will also be valid for the other general strategy for selection from among nested models namely, RJMCMC.

Thank You !