Pólya-gamma data augmentation and latent variable models for multivariate binomial data: Other results

John B. Holmes
University of Melbourne, Melbourne, Australia.
E-mail: john.holmes@unimelb.edu.au
Matthew R. Schofield
University of Otago, Dunedin, New Zealand.
Richard J. Barker
University of Otago, Dunedin, New Zealand.

1. Implied priors under identification constraints

In the main paper, we provided a method to obtain identifiable estimates of $\mathbf{F}$ and $\mathbf{\Lambda}$ by applying SVD to unconstrained estimates at each iteration of the Gibbs sampler. What we have not shown is the impact these constraints have on the implied priors for $\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{F}}, \tilde{\mathbf{\Lambda}}$ and $\tilde{\boldsymbol{\sigma}}^2_\lambda$. To determine the implied priors, we simulated 10,000 sets of $\mathbf{F}, \mathbf{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2_\lambda$ using the prior specifications/initial values, $\mathbf{F}_i \sim \mathcal{N}(0, \mathbf{I}_f) \quad (i = 1, \ldots, n)$, $\mathbf{\Lambda}_l \sim \mathcal{N}(0, \sigma^2_{\lambda_l}\mathbf{I}_k) \quad (l = 1, \ldots, f)$, $\boldsymbol{\mu} = 0$, $\sigma^2_{\lambda_l} \sim IG(\alpha_{\mathbf{\Lambda}}, \beta_{\mathbf{\Lambda}})$ where $\alpha_{\mathbf{\Lambda}} = 2.1$, $\beta_{\mathbf{\Lambda}} = 1.1\pi^2/3f$ and $f = 4, k = 10, n = 80$. We then applied column centring and SVD to $\mathbf{1}_n\boldsymbol{\mu}' + \mathbf{F}\mathbf{\Lambda}'$. For $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{\Lambda}}$, the prior is the same as that of $\mathbf{F}$ and $\mathbf{\Lambda}$ (see Fig. 1), $\tilde{\mathbf{F}}_i \sim \mathcal{N}(0, \mathbf{I}_f) \quad (i = 1, \ldots, n)$ and $\tilde{\mathbf{\Lambda}}_l \sim \mathcal{N}(0, \tilde{\sigma}^2_{\lambda_l}\mathbf{I}_k) \quad (l = 1, \ldots, f)$. The difference compared to $\mathbf{F}, \mathbf{\Lambda}$ is that the orthogonality constraint in SVD strictly enforces independence of the columns of $\tilde{\mathbf{F}}$, and $\tilde{\mathbf{\Lambda}}$.

Differences between the original specification/prior and implied prior distribution appear in $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\sigma}}^2_\lambda$. We initially fix $\boldsymbol{\mu}$ to zero and assume the prior is flat, however the implied prior on $\boldsymbol{\mu}$ is the mean of n product-normal distributions, each with zero mean and variance $\sum_{l=1}^{f} \sigma^2_{\lambda_l}$, which with our choice of $\alpha_{\mathbf{\Lambda}}, \beta_{\mathbf{\Lambda}}$ has an expected value of $\pi^2/3$. This means the prior for $\tilde{\boldsymbol{\mu}}$ has heavier tails than a $\mathcal{N}(0, \pi^2/3n)$ prior (Fig. 2). The most pronounced change is in the implied prior for $\tilde{\boldsymbol{\sigma}}^2_\lambda$. Performing SVD is equivalent to sorting $\boldsymbol{\sigma}^2_\lambda$, so the implied prior for $\tilde{\boldsymbol{\sigma}}^2_\lambda$ does not assume equally important factors like the prior for $\boldsymbol{\sigma}^2_\lambda$. While it may appear that the sorting is equivalent to placing a truncated inverse-gamma prior on $\tilde{\sigma}^2_{\lambda_l}$, this is not the case. In Figure 2, we compare the implied prior for $\tilde{\sigma}^2_{\lambda_l}$ to $IG(\alpha_{\mathbf{\Lambda}}, \beta_{\mathbf{\Lambda}})$ truncated to $[a, b]$ where a, b are the $(l-1)/f$ and l/f quantiles of $IG(\alpha_{\mathbf{\Lambda}}, \beta_{\mathbf{\Lambda}})$. Compared to the truncated distribution, the implied prior for $\tilde{\sigma}^2_{\lambda_l}$ is shifted left, especially as the index l increases and has a longer right tail. This means the implied prior for $\tilde{\sigma}^2_{\lambda_l}$ exhibits more rapid decay than a standard inverse gamma prior.

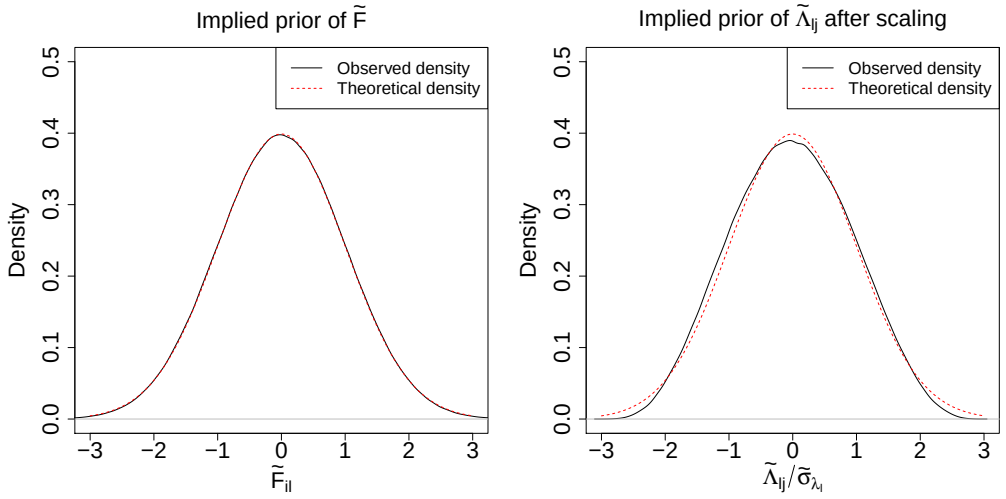

Fig. 1. The implied prior distribution of $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{\Lambda}}$ divided by $\tilde{\boldsymbol{\sigma}}^2_{\lambda}$ obtained from 10,000 simulated sets of $\mathbf{F}, \mathbf{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2_{\lambda}$.

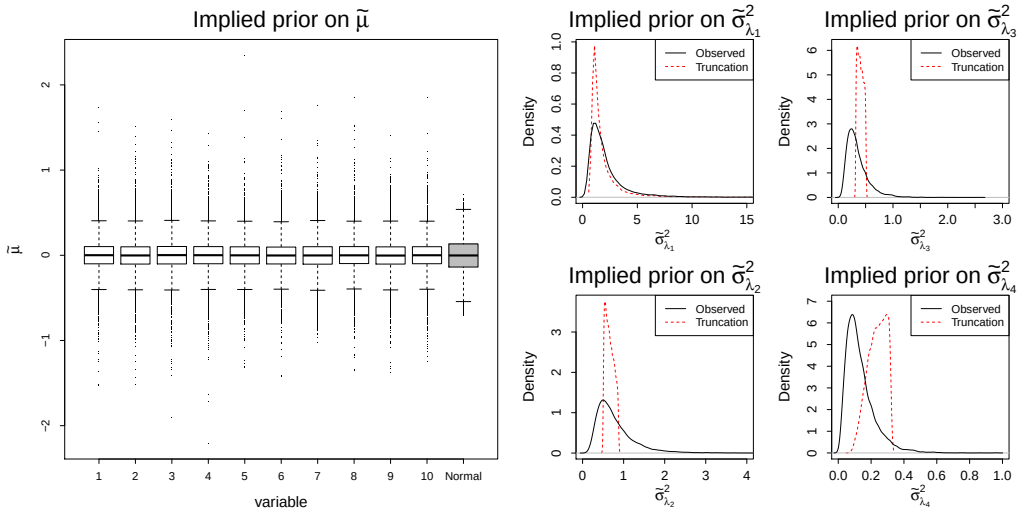

Fig. 2. The implied priors of $\tilde{\boldsymbol{\mu}}$ (left) and $\tilde{\boldsymbol{\sigma}}^2_{\lambda}$ (right panel) obtained from 10,000 simulated sets of $\mathbf{F}, \mathbf{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2_{\lambda}$.The grey boxplot with the boxplots of the implied prior of $\boldsymbol{\mu}$ corresponds to the distribution $\mathcal{N}(0, \pi^2/3n)$. The dashed line on the implied prior of $\tilde{\boldsymbol{\sigma}}^2_{\lambda}$ corresponds to a truncation of the inverse-gamma prior assumed for $\boldsymbol{\sigma}^2_{\lambda}$.

2. $\mathbf{N}$ as an additional layer of replication

In the main text, we discussed how the binomial trial size, $\mathbf{N}$, is an additional layer of replication in the multivariate binomial dataset $\mathbf{Y}$, and that the posterior mean of $\boldsymbol{\theta}$ is consistent in respect to $\mathbf{N}$. If true, we would then expect the accuracy of parameter estimates to improve as $\mathbf{N}$ increases. We then showed this was the case for $\tilde{\boldsymbol{\sigma}}^2_{\lambda}$. Here,

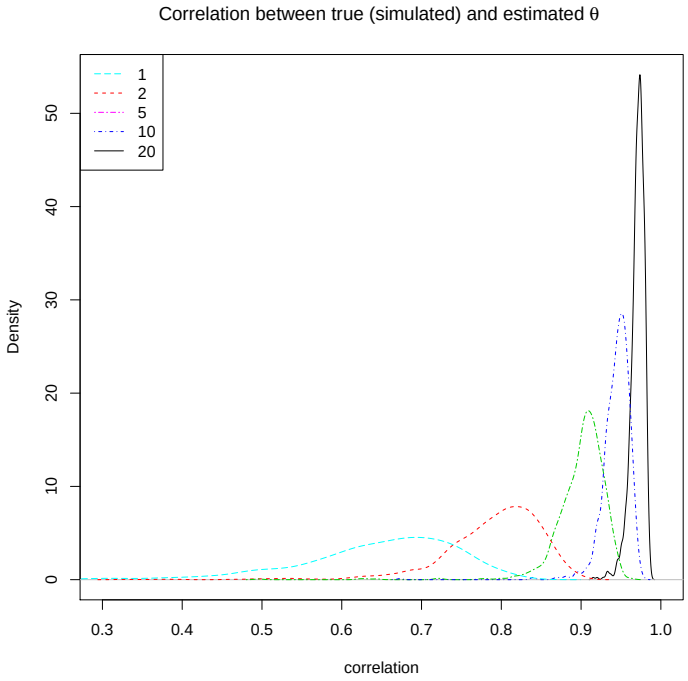

Fig. 3. The distribution of correlation of true (simulated) and posterior mean estimate of $\boldsymbol{\theta}_j$ obtained from the blocked Gibbs sampler of the 100 replicate datasets over a range of binomial trial sizes from the four dimensional model.

we check if this is true for $\boldsymbol{\theta}$ by examining Pearson correlation between the simulated realisations of the natural parameter for observed variable j, $\boldsymbol{\theta}_j$ and the posterior mean estimate of $\boldsymbol{\theta}_j$ obtained from the four dimensional model for $N_{ij} = 1, 2, 5, 10, 20\ \forall i, j$. In Figure 3, we plot the distribution of correlation by trial size over all observed variables. This confirms increased $\mathbf{N}$ does lead to improved estimates of $\boldsymbol{\theta}$. This matches with the inverse relationship between shrinkage to $\boldsymbol{\mu}$ and $\mathbf{N}$ in Corollary 1 of the main paper.

3. Properties of our approach when elements of σ^2_λ are poorly separated

While our post-hoc transformations of $\mathbf{A}$ or $\boldsymbol{\theta}$ were designed to prevent rotation, and produce identified estimates $\tilde{F}, \tilde{\boldsymbol{\Lambda}}, \tilde{\boldsymbol{\sigma}}^2_\lambda$, we need to distinguish between the case when elements of $\tilde{\boldsymbol{\sigma}}^2_\lambda$ are distinct, as considered in Section 8, and when some or all elements of $\tilde{\boldsymbol{\sigma}}^2_\lambda$ are equal. When elements of $\tilde{\boldsymbol{\sigma}}^2_\lambda$ are 'close' together, the random variation in the link matrix $\boldsymbol{\theta}$ could mean the relative importance of a factor swaps between iterations of the Gibbs sampler. To determine if our method to obtain identification of estimates can cope with this, we performed 60 simulations with the loading variance on factors 2 and 3 tied, such that $\sigma^2_{\lambda_1} = 1.50$, $\sigma^2_{\lambda_2} = \sigma^2_{\lambda_3} = 0.97$, $\sigma^2_{\lambda_4} = 0.61$, with $\mathbf{Y}$ generated with trial size $\mathbf{N} = 20$ for all elements and $\boldsymbol{\theta}$ generated using a principal component representation. To quantify performance, we focused on three measures when fitting PG-PCA:

Gelman-Rubin statistics, to determine whether MCMC chains converged for $\tilde{F}, \tilde{\mathbf{\Lambda}}, \tilde{\boldsymbol{\sigma}}^2_\lambda$, the posterior distribution of $\tilde{\boldsymbol{\sigma}}^2_{\lambda l}$, to determine whether distributions were overlapping for $\tilde{\sigma}^2_{\lambda_2}, \tilde{\sigma}^2_{\lambda_3}$ and uni-modal, and the posterior distributions of the elements of $\tilde{\mathbf{\Lambda}}$, to check for multi-modality, especially in factors 2 and 3.

We fitted PG-PCA to 60 simulated datasets running a Gibbs sampler for 15,000 iterations, discarding the first 5,000 iterations with three chains. For all elements of $\tilde{\boldsymbol{\mu}}$, the convergence diagnostic was between 1.000 and 1.007 for all simulations. For the elements of $\tilde{\boldsymbol{\sigma}}^2_\lambda$, the convergence diagnostic was between 1.000 and 1.009 in all simulations except one, where the convergence diagnostic for one element of $\tilde{\boldsymbol{\sigma}}^2_\lambda$ was 1.016. For the elements of $\tilde{\mathbf{\Lambda}}$, the convergence diagnostic was between 1.000 and 1.009 in all simulations except two, where the convergence diagnostic for one element of $\tilde{\mathbf{\Lambda}}$ was 1.010 and 1.011 respectively. For the elements of $\tilde{\mathbf{F}}$, the convergence diagnostic was between 1.001 and 1.009 for all simulations, except in one simulation where one element had a convergence diagnostic of 1.011. Having Gelman-Rubin diagnostics near one indicates that our approach to enforcing identification ensures is not leading to sampling in distinct sub-spaces in the different chains, but does not prove posterior distributions are not multi-modal.

In Figures 4 and 5, density plots of the posterior distributions of the elements of $\tilde{\boldsymbol{\sigma}}^2_\lambda$ and of $\tilde{\sigma}^2_{\lambda_2} - \tilde{\sigma}^2_{\lambda_3}$ for the first 20 simulations are shown. For most simulations, the posterior $[\tilde{\boldsymbol{\sigma}}^2_\lambda | \mathbf{Y}]$ is uni-modal for all elements, which is suggestive of low levels of label switching and/or rotation between iterations. However some simulations do show multi-modality in the posteriors, but can be seen in Figure 4 for simulations 6, 11, and 17 this multi-modality is not always between factors 2 and 3. While our approach to enforcing identification, outlined in section 7.3 of the main paper, allows for the relative importance of factors to switch, the posterior distribution of $\tilde{\sigma}^2_{\lambda_2} - \tilde{\sigma}^2_{\lambda_3}$, the difference of the two loading variances we simulated to be equal, is bi-modal with modes equally distant from zero, not uni-modal with the mode at zero. We believe this makes intuitive sense as the SVD algorithm aims to rank factors so the randomness in each realisation of $\tilde{\boldsymbol{\theta}}$ ensures a certain level of separation between the estimates $\tilde{\sigma}^2_{\lambda_2}, \tilde{\sigma}^2_{\lambda_3}$ in each iteration of the Gibbs sampler.

In Figure 6-7, density plots of the posterior distributions of the elements of $\tilde{\mathbf{\Lambda}}$ are shown from two simulated datasets, 4 and 17. In Figure 6, which displays the posterior of the elements of $\tilde{\mathbf{\Lambda}}$ from simulation 4, where multi-modality was not present in the posterior of $\tilde{\boldsymbol{\sigma}}^2_\lambda$, we only observe evidence of multi-modality in factors 2 and 3, where $\sigma^2_{\lambda_l}$ was simulated to be equal. This multi-modality was picked up by all chains of the Gibbs sampler. In Figure 7, which displays the posterior of the elements of $\tilde{\mathbf{\Lambda}}$ from simulation 17, where multi-modality was present in the posterior of $\tilde{\boldsymbol{\sigma}}^2_\lambda$ for factors 2 and 4, we observe evidence of multi-modality in factors 2, 3 and 4, so for both the elements where $\sigma^2_{\lambda_l}$ was simulated to be equal, and the factors where multi-modality in the posterior of $\sigma^2_{\lambda_l}$ was observed. As with simulation 4, this multi-modality was picked up by all chains of the Gibbs sampler. While it would be instinctive to believe these results are just indicative of label switching, we believe there is also partial rotation occurring when elements of $\tilde{\boldsymbol{\sigma}}^2_\lambda$ are similar in value, as our approach to enforcing identification relies on the correlation of singular factors from the SVD of $\mathbf{A}$ and $\mathbf{Y}$, so will struggle to identify the right factor in Gibbs sampler iterations where $|\text{Cor}(\mathbf{V}_l, \mathbf{V}_{Y_l})|, |\text{Cor}(\mathbf{V}_{l'}, \mathbf{V}_{Y_{l'}})|, |\text{Cor}(\mathbf{V}_{l'}, \mathbf{V}_{Y_l})|$, $|\text{Cor}(\mathbf{V}_l, \mathbf{V}_{Y_{l'}})|$ are similar.

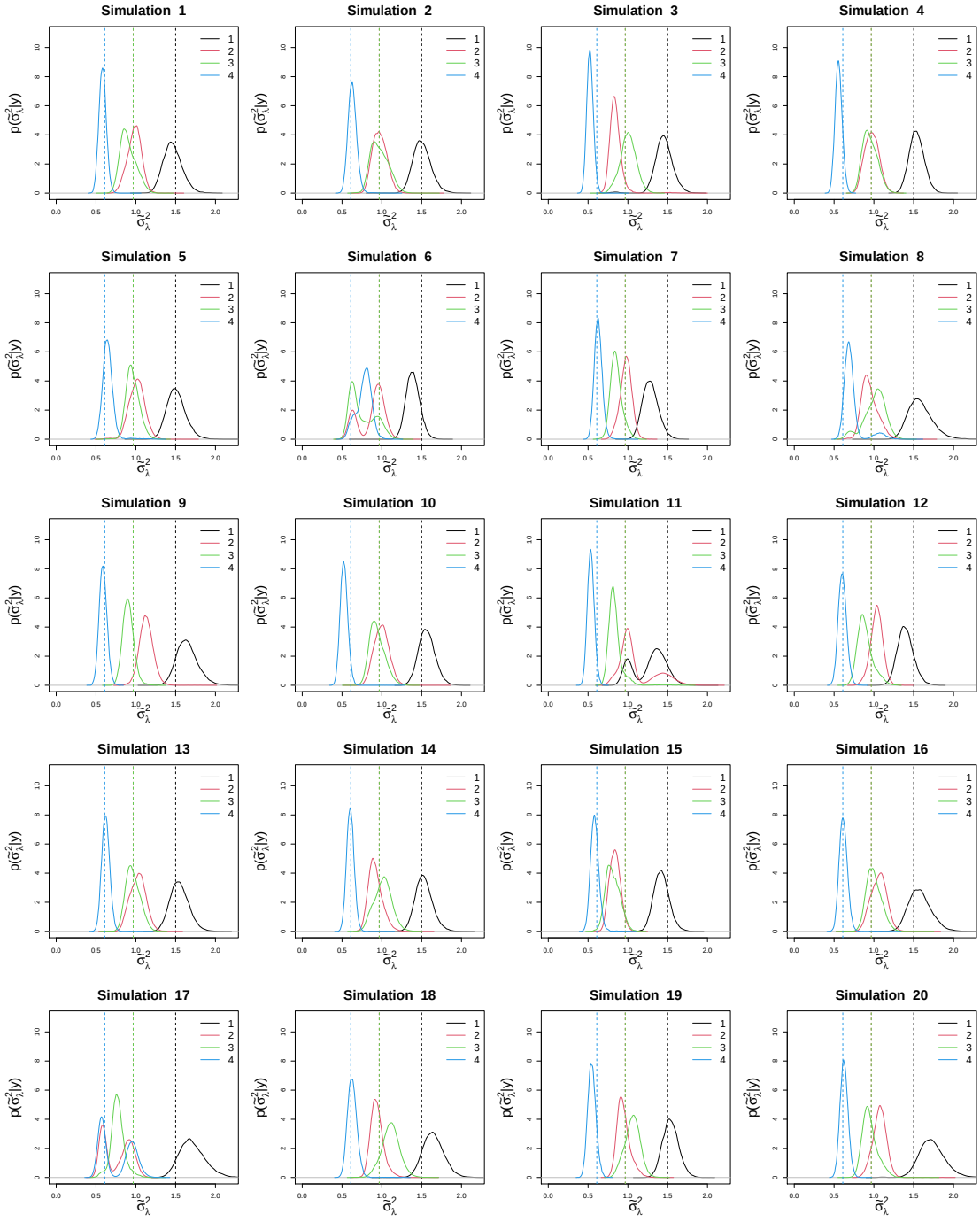

Fig. 4. The posterior distribution of $\tilde{\sigma}^2_\lambda$ obtained from the blocked Gibbs sampler for 20 replicate datasets from the four dimensional model with $\sigma^2_{\lambda_2} = \sigma^2_{\lambda_3}$ in the simulation. Vertical lines indicate the simulated values for the elements of $\boldsymbol{\sigma}^2_\lambda$.

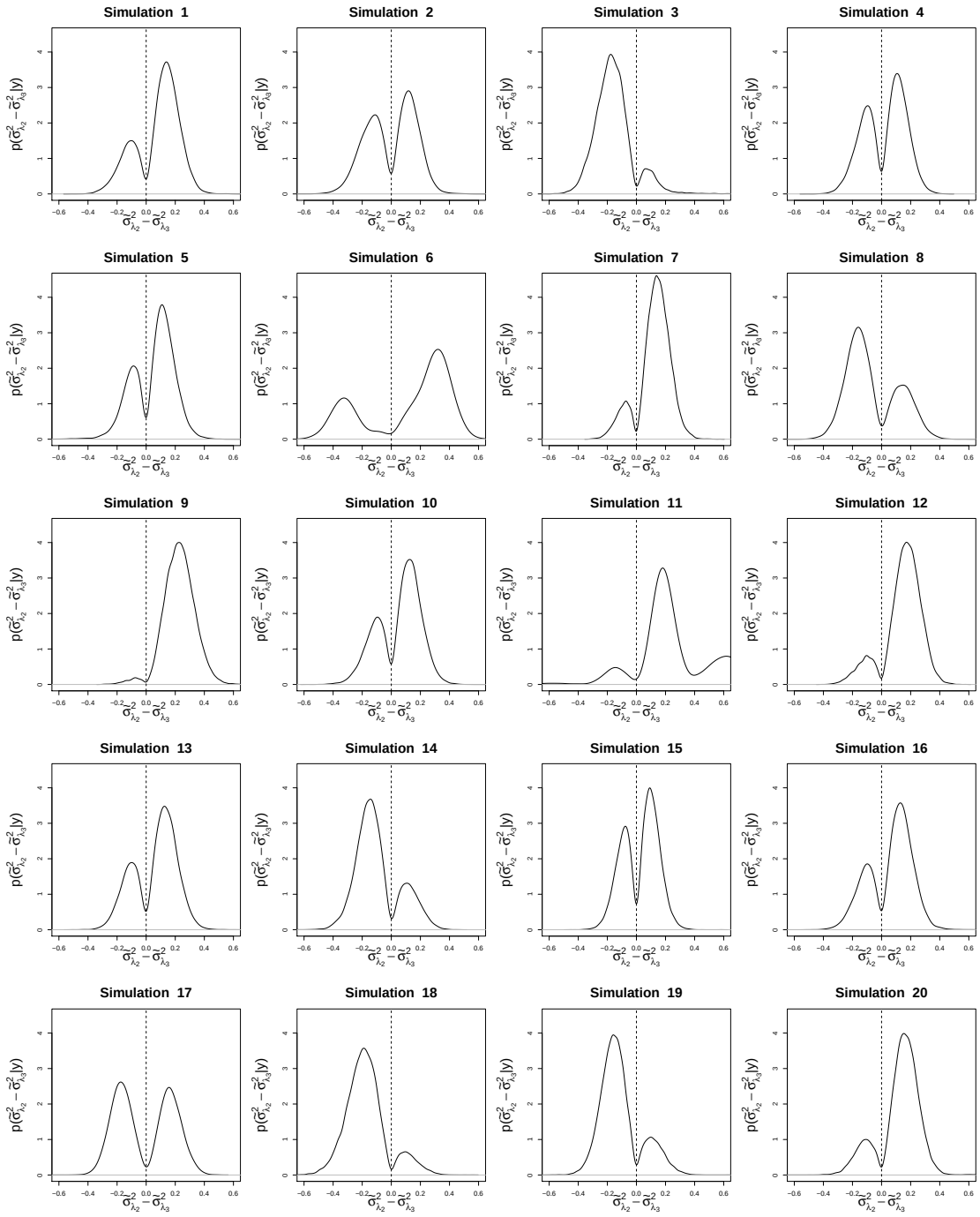

Fig. 5. The posterior distribution of $\tilde{\sigma}^2_{\lambda_2} - \tilde{\sigma}^2_{\lambda_3}$ obtained from the blocked Gibbs sampler for 20 replicate datasets from the four dimensional model with $\sigma^2_{\lambda_2} = \sigma^2_{\lambda_3}$ in the simulation.

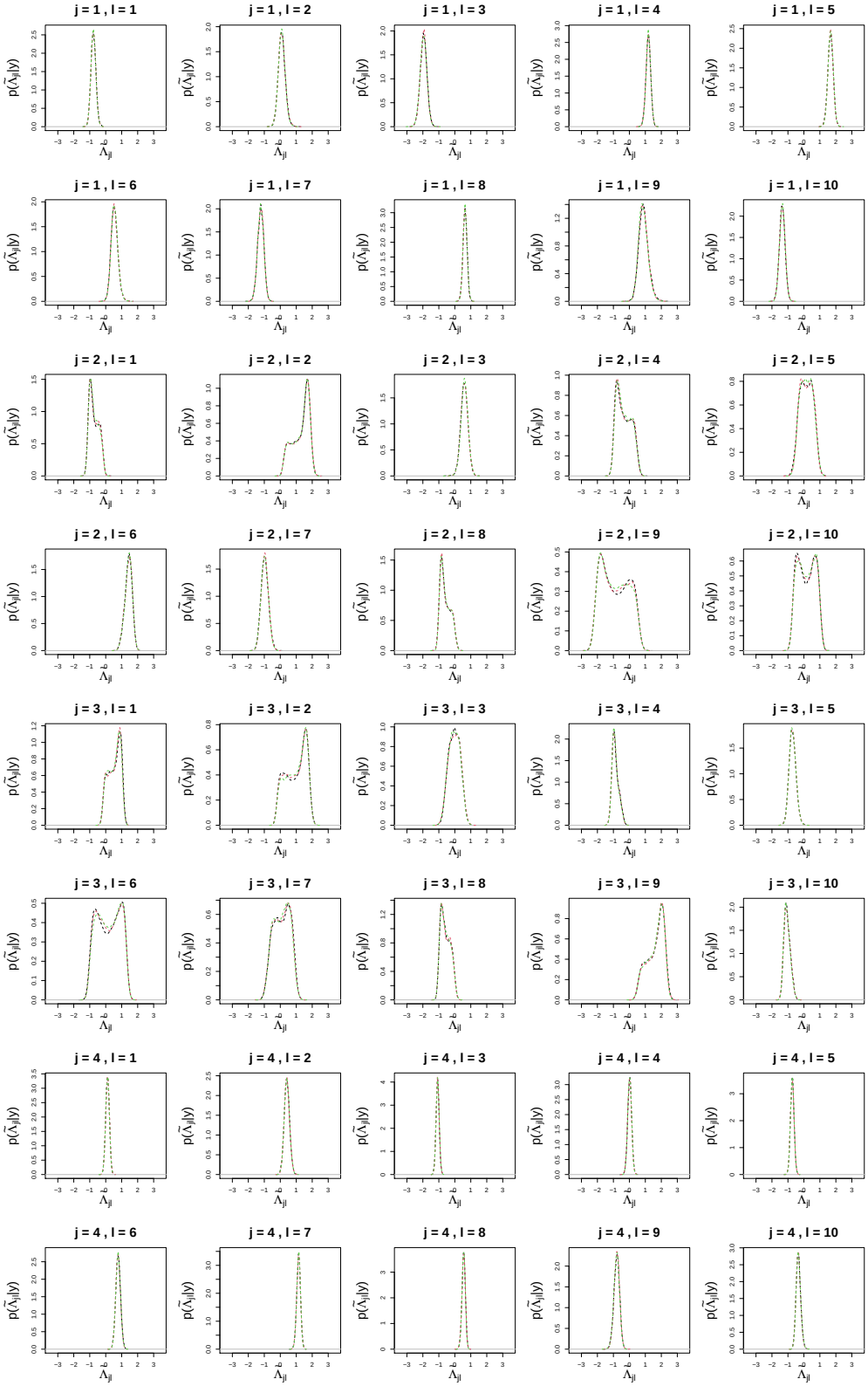

Fig. 6. The posterior distribution of the elements $\tilde{\Lambda}_{jl}$ obtained from the blocked Gibbs sampler analysing simulated dataset 4 from the four dimensional principal component model. j is the observed variable, l is the factor. Black, red and green dashed lines indicates different chains.

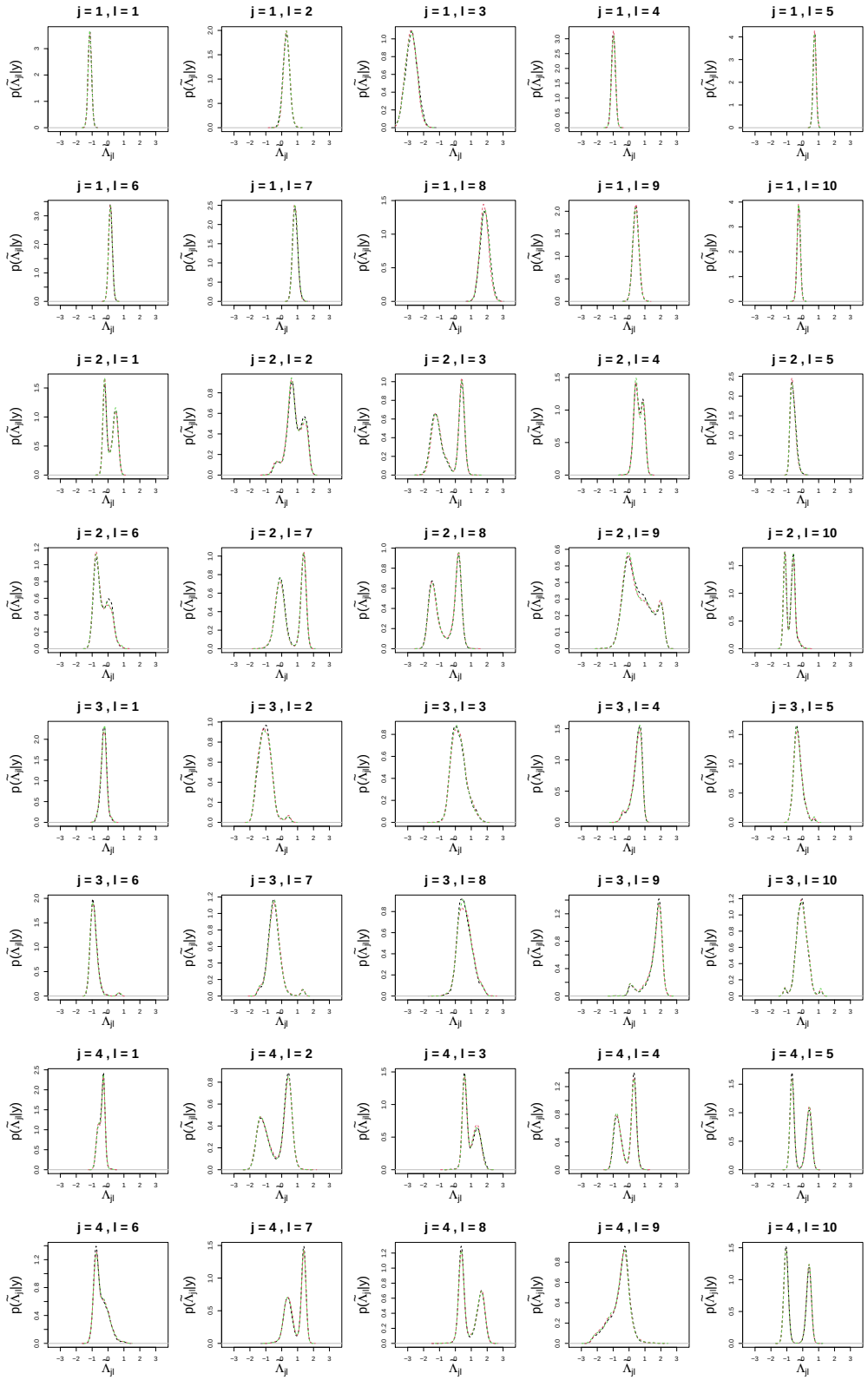

Fig. 7. The posterior distribution of the elements $\tilde{\Lambda}_{jl}$ obtained from the blocked Gibbs sampler analysing simulated dataset 17 from the four dimensional principal component model. j is the observed variable, l is the factor. Black, red and green dashed lines indicates different chains.

Table 1. Posterior estimated loadings and standard deviation for the two factors when fitting the principal component model to the New Zealand apprehension and prosecution data when $\mathbf{\Lambda}_M \neq \mathbf{\Lambda}_{NM}$ (first eight columns) and $\mathbf{\Lambda}_M = \mathbf{\Lambda}_{NM}$ (last four columns).

Observed variable	Factor 1, Māori		Factor 1, non-Māori		Factor 2, Māori		Factor 2, non-Māori		Factor 1, $\mathbf{\Lambda}_M = \mathbf{\Lambda}_{NM}$		Factor 2, $\mathbf{\Lambda}_M = \mathbf{\Lambda}_{NM}$	
Homicide	-0.132	0.198	0.298	0.109	0.369	0.152	-0.079	0.099	0.123	0.073	0.074	0.075
Intent to cause injury	0.063	0.014	0.057	0.016	-0.065	0.025	-0.074	0.025	0.042	0.007	-0.087	0.004
Sexual assault	0.103	0.057	0.177	0.077	0.150	0.035	0.206	0.041	0.211	0.018	0.153	0.019
Dangerous/ negligence	0.308	0.059	0.287	0.052	-0.074	0.087	-0.064	0.081	0.280	0.023	-0.147	0.024
Abduction/ harassment	0.209	0.045	0.206	0.055	0.035	0.049	0.090	0.044	0.228	0.006	0.018	0.018
Robbery	0.290	0.124	0.265	0.112	0.331	0.066	0.286	0.061	0.366	0.026	0.246	0.027
Burglary	0.238	0.087	0.253	0.103	0.215	0.050	0.275	0.055	0.309	0.014	0.178	0.024
Theft	0.211	0.058	0.193	0.043	0.103	0.045	0.045	0.044	0.223	0.003	0.022	0.017
Fraud	0.127	0.061	-0.040	0.016	0.171	0.034	0.065	0.022	0.034	0.010	0.102	0.009
Illicit drugs	0.190	0.032	0.224	0.037	-0.033	0.051	-0.044	0.061	0.203	0.008	-0.096	0.015
Weapons/ explosives	0.238	0.056	0.261	0.066	0.065	0.054	0.098	0.057	0.277	0.007	0.021	0.021
Property/ Environment	0.112	0.038	0.145	0.046	0.088	0.024	0.101	0.030	0.158	0.006	0.063	0.012
Public order	0.284	0.083	0.273	0.101	-0.398	0.129	-0.480	0.141	0.163	0.040	-0.524	0.012
Justice/ government	0.212	0.040	0.308	0.073	0.000	0.053	0.090	0.069	0.291	0.005	-0.014	0.022
Miscellaneous	0.331	0.063	0.243	0.048	-0.231	0.114	-0.204	0.089	0.216	0.027	-0.283	0.019

For each heading, first column is posterior mean, second is posterior standard deviation.

4. Estimated factor loadings $\tilde{\Lambda}_{jl}$ using the various techniques

To help interested readers gain an appreciation of the patterns found when fitting PG-PCA and PG-FA to the police apprehension and prosecution records, we present Tables 1-2. These tables contain information on the estimated factor loadings $\tilde{\mathbf{\Lambda}}$ for both when we enforce the constraint $\mathbf{\Lambda}_M = \mathbf{\Lambda}_{NM}$ or when we do not. There is one feature we wish to discuss. The level of uncertainty in estimated loadings $\tilde{\Lambda}_{jl}$ is lower when fitting PG-PCA compared to PG-FA. This is unsurprising as the inclusion of an error component in $\boldsymbol{\theta}$ means more parameters are estimated in PG-FA.

Table 2. Posterior estimated loadings and standard deviation for the two factors when fitting the factor analysis model to the New Zealand apprehension and prosecution data when $\mathbf{\Lambda}_M \neq \mathbf{\Lambda}_{NM}$ (first eight columns) and $\mathbf{\Lambda}_M = \mathbf{\Lambda}_{NM}$ (last four columns). For each heading, first column is posterior mean, second is posterior standard deviation.

Observed variable	Factor 1, Māori		Factor 1, non-Māori		Factor 2, Māori		Factor 2, non-Māori		Factor 1, $\mathbf{\Lambda}_M = \mathbf{\Lambda}_{NM}$		Factor 2, $\mathbf{\Lambda}_M = \mathbf{\Lambda}_{NM}$	
Homicide	0.028	0.223	0.260	0.138	0.225	0.173	-0.105	0.120	0.202	0.112	-0.003	0.097
Intent to cause injury	0.052	0.032	0.043	0.030	-0.072	0.037	-0.079	0.035	0.045	0.021	-0.082	0.018
Sexual assault	0.113	0.074	0.203	0.088	0.134	0.057	0.175	0.061	0.181	0.044	0.161	0.039
Dangerous/ negligence	0.289	0.081	0.274	0.072	-0.114	0.095	-0.101	0.088	0.291	0.043	-0.127	0.048
Abduction/ harassment	0.212	0.053	0.215	0.060	0.003	0.056	0.054	0.053	0.224	0.019	0.020	0.036
Robbery	0.329	0.130	0.306	0.121	0.278	0.082	0.240	0.079	0.348	0.055	0.265	0.052
Burglary	0.263	0.090	0.284	0.108	0.173	0.062	0.236	0.070	0.292	0.038	0.197	0.045
Theft	0.220	0.062	0.194	0.051	0.069	0.052	0.016	0.051	0.218	0.019	0.037	0.035
Fraud	0.148	0.073	-0.006	0.053	0.136	0.055	0.055	0.052	0.090	0.044	0.103	0.037
Illicit drugs	0.182	0.043	0.210	0.047	-0.063	0.057	-0.077	0.065	0.202	0.021	-0.083	0.033
Weapons/ explosives	0.244	0.065	0.269	0.073	0.027	0.063	0.060	0.065	0.271	0.023	0.035	0.042
Property/ Environment	0.124	0.044	0.153	0.050	0.068	0.034	0.073	0.040	0.147	0.019	0.068	0.026
Public order	0.217	0.100	0.194	0.118	-0.423	0.135	-0.494	0.146	0.196	0.079	-0.486	0.035
Justice/ government	0.200	0.049	0.304	0.077	-0.031	0.058	0.035	0.074	0.255	0.027	-0.013	0.041
Miscellaneous	0.256	0.103	0.226	0.084	-0.272	0.120	-0.194	0.099	0.245	0.068	-0.257	0.049

For each heading, first column is posterior mean, second is posterior standard deviation.