Multivariate Modelling: Latent Variables & Dimensionality Reduction

Habib Ganjgahi

Department of Statistics
University of Oxford

November 6, 2018

Table of Contents

Introduction to Neuroimaging

Go to Power point

Table of Contents

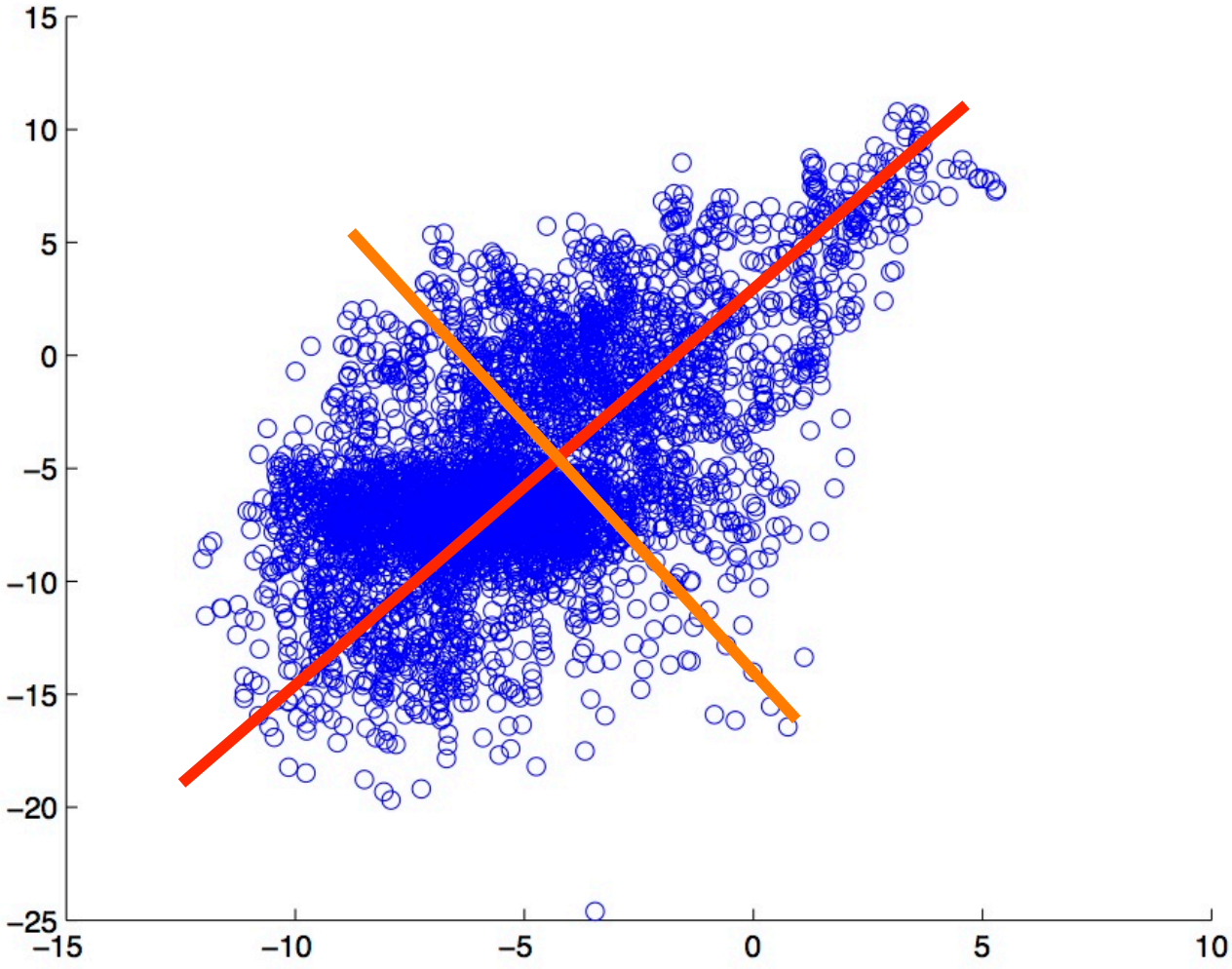
Table of Contents

Latent Variable Modelling

- Model high dimensional Y with a lower-dimensional approximation F that has N rows, but typically far fewer columns
- Latent variable modelling is used for:
  - Dimensionality reduction: replace Y with a lower-dimensional F
  - Outlier detection
  - Data visualization: display F in a scatterplot.
  - Factor discovering: discover important hidden **factors** underlying data.

PCA

Projects the data into a sub-space such that the projected data has maximum variance.



Principal Component Analysis (PCA)

- Projects the data into a sub-space such that the projected data has maximum variance.

$$\Upsilon \approx F\Lambda',$$

where $\Upsilon \in \mathbb{R}^{N\times P}$; $\Lambda \in \mathbb{R}^{P\times K}$ principal axis; $F \in \mathbb{R}^{N\times K}$ and $K << P$. PCA derivations:

- choose Λ minimizes construction error: $\text{argmin}\|\Upsilon - F\Lambda'\|_{\text{F}}^2$
- choose Λ maximises the variance: $\text{argmax}\sum_{i=1}^{N}\|\Lambda'\Upsilon_i\|^2 = \text{Tr}(\Lambda'\Upsilon'\Upsilon\Lambda)$
- Λ is the eigenvetor of sample covariance matrix $S = \frac{1}{N}\Upsilon'\Upsilon$

Table of Contents

Factor Analysis (FA)

Describes the correlation among variables using few latent variables. Projects the data into a space that are not correlated

$$\boldsymbol{y}_i = \boldsymbol{f}_i \boldsymbol{\Lambda}' + \boldsymbol{\epsilon}_i, \tag{1}$$

where $k \ll N$ is; $\boldsymbol{f}_t \in \mathbb{R}^K$ is the k vector of latent variables (scores); and $\boldsymbol{\Lambda} \in \mathbb{R}^{P \times k}$ is the loadings matrix. It is assumed that

- Latent variables are independent $\boldsymbol{f}_i \sim N(\boldsymbol{0}, \boldsymbol{I}_{k \times k})$
- Residuals are independent with $\boldsymbol{\epsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{\Psi})$ where $\boldsymbol{\Psi} = \text{diag}(\sigma_1^2, \cdots, \sigma_P^2)$
- The latent variables and residuals are mutually independent $\text{Cov}(\boldsymbol{f}_t, \boldsymbol{\epsilon}_s) = 0$ for all t and s.

Factor Analysis

The FA model:

$$\boldsymbol{y}_i = \boldsymbol{f}_i \boldsymbol{\Lambda}' + \boldsymbol{\epsilon}_i,$$

- $\boldsymbol{y}_i | \boldsymbol{f}_i \sim \mathcal{N}(\boldsymbol{f}_i \Lambda', \boldsymbol{\Psi})$
- $\boldsymbol{y_i} \sim \mathcal{N}(0, \Sigma)$, $\Sigma = \Lambda\Lambda' + \boldsymbol{\Psi}$
- log-likelihood function

$$l(\Lambda, \boldsymbol{\Psi}) \propto -\frac{N}{2} \log |\boldsymbol{\Psi}| - \frac{N}{2} \mathrm{Tr}(S\boldsymbol{\Psi}^{-1})$$

  where $S = \frac{1}{N} \boldsymbol{Y}' \boldsymbol{Y}$ is the sample covariance matrix
- Parameter estimation: Maximise likelihood function
- Varimax rotation for interpretation

Dimensionality Reduction: Unified Approach

$$
\begin{aligned}
\boldsymbol{Y} &= \boldsymbol{F}\Lambda' + \boldsymbol{E}, \\
\boldsymbol{F} &\sim \mathcal{N}(\boldsymbol{0}, I), \quad E \sim \mathcal{N}(\boldsymbol{0}, \Psi), \\
\boldsymbol{Y} &\sim \mathcal{N}(\boldsymbol{0}, \Lambda\Lambda' + \Psi), \\
\boldsymbol{F}|\boldsymbol{Y}, \Lambda, \Psi &\sim \mathcal{N}(m, V) \\
V &= \left(I - \Lambda'\Psi^{-1}\Lambda\right)^{-1}, \quad m = V\Lambda'\Psi^{-1}\boldsymbol{Y}
\end{aligned}
$$

- PCA [1]
  - $\Psi = \sigma^2 I$
  - $\sigma^2 \to 0$ recovers classic PCA
- FA
  - $\Psi = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_P^2)$

Identifiability

The model (Eq., 1) is invariant under any orthogonal transformation:

$$\boldsymbol{y} = (\boldsymbol{F}P)(\boldsymbol{\Lambda}P)' + \boldsymbol{\epsilon} = \boldsymbol{F}^*\boldsymbol{\Lambda}^* + \boldsymbol{\epsilon}$$

where $PP' = I$. That is mean that the rotated parameters $(\boldsymbol{F}^*, \boldsymbol{\Lambda}^*)$ and $(\boldsymbol{F}, \boldsymbol{\Lambda})$ are not distinguishable, hence further restrictions, so called identifiability condition, should be imposed to achieve unique solution.

Identifiability

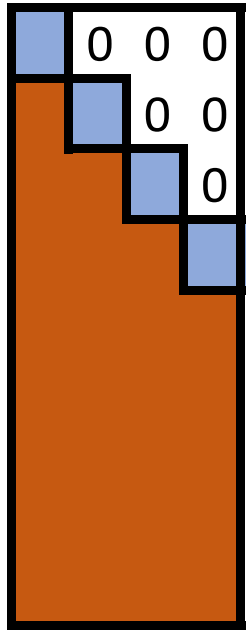
- Constraint parameter space
  - Block of lower triangular matrix
  - Sparse mode
- Post process MCMC outputs or MAP estimates

Table of Contents

Bayesian FA: Identifiability

Identifiability condition breaks the symmetry of likelihood function and guarantees unique mode of the posterior[2]:

- Λ is full column rank and has a square block of lower triangular with positive diagonal elements (LT).
- $\exists B$ of size Λ and full column rank such that $B'\Lambda$ has a block of lower triangular matrix.
- $\Lambda'\Lambda$ or $\Lambda'\Sigma^{-1}\Lambda$ is diagonal and diagonal elements are unique and arranged in descending order.



Inference

- Bayes rule:

$$\pi(\theta|Y) \propto \pi(\theta)\pi(Y|\theta),$$

  where $\theta = (\Lambda, \boldsymbol{F}, \Psi)$
- MCMC algorithm
- MAP estimates: Expectation Maximisation (EM)

$$\begin{aligned} Q(\theta) &= E_{\boldsymbol{F}|.} \log(\pi(\theta|Y)) \\ \hat{\theta} &= \operatorname{argmax}(Q(\theta)) \end{aligned}$$

Identifiability: LT

- Caveats:
  - Posterior may suffer from local maxima that affects the MCMC sampling schemes.
  - Reduces the accuracy of parameter estimations [3].
  - Ordering problem: LT constraint is invariant to order of variables
  - Number of latent variables is influenced by order of variables [4].

Identifiability

- Identifiability condition that $\mathbf{\Lambda}\mathbf{\Lambda}'$ is diagonal with unique diagonal elements in descending order.
- Kaufmann and Schumacher[5] used the following specification for the FA model in Eq.,(1)

$$\begin{aligned} \boldsymbol{y}_t &= (\boldsymbol{f}_t^* P)(\mathbf{\Lambda}^* P)' + \boldsymbol{\epsilon}, \\ &= \boldsymbol{f}_t \mathbf{\Lambda}' + \boldsymbol{\epsilon}, \end{aligned}$$

  where $\mathbf{\Lambda}^*$ is sparse; $\boldsymbol{f}_t^*$ is the latent variable vector corresponds to the sparse loading and P is an orthogonal matrix such that $\mathbf{\Lambda}'\mathbf{\Lambda}$ is diagonal with unique elements in descending order.
- P is derived from eigenvalue decomposition of $\mathbf{\Lambda}^*\mathbf{\Lambda}^*$.
- In their approach, a draw from $\mathbf{\Lambda}^*$ is used to get an identified solution, then latent variables correspond to the sparse loading is derived using the identified latent variable as $\boldsymbol{f}_t^* = \boldsymbol{f}_t P'$.

Identifiability: Sparse Modes

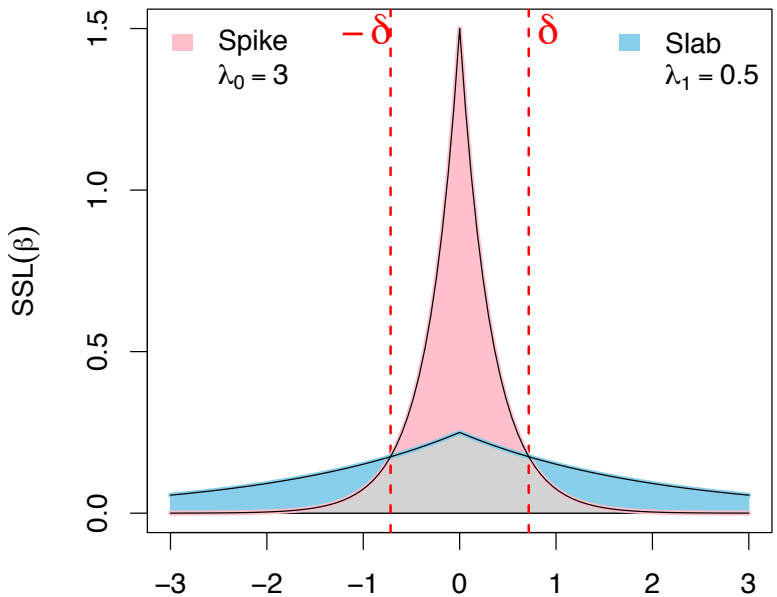
- Sparsity inducing priors can be used to achieve unique solution
- Helps interoperability of latent variables while only few variables enter to each latent variables.
- Improves predictive performances [6].
- Shrinkage prior:
  - Automatic relevance determination (ARD)
  - Horseshoe
- No exact zeros
- spike and slap prior

Spike and Slab prior

Spike and Slab (SSL) prior with Laplace components

$$\pi(\lambda|\gamma) = \gamma \text{Laplace}(\lambda|\sigma_1^2) + (1-\gamma)\text{Laplace}(\lambda|\sigma_0^2)$$

- σ_1^2 **small**: to avoid over-shrinkage of large effects
- σ_0^2 **large**: to shrink ignorable coefficients to zero
- θ controls sparsity, $P(\gamma = 1|\theta) = \theta$
- Point-mass spike-and-slab is a limiting case as $\sigma_0^2 \to 0$



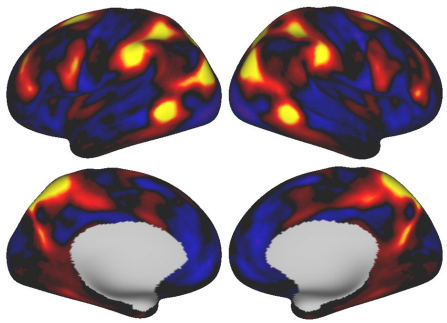
Latent variable: Spike and Slab prior

Hierarchical spike and slab prior[7]:

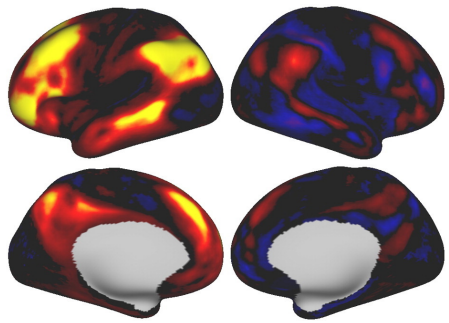
$$
\begin{aligned}
\pi(\lambda_{jk}|\gamma_{jk}) &\sim \gamma \text{Laplace}(\lambda_{jk}|\sigma_1^2) + (1-\gamma)\text{Laplace}(\lambda_{jk}|\sigma_0^2) \\
\gamma_{jk} &\sim \text{Bernoulli}(\theta_k) \\
\theta_k &= \prod_{l=1}^{k} \nu_l \\
\nu_l &\sim^{iid} \text{B}(\alpha, 1)
\end{aligned}
$$

An Example: rsfMRI Modelling[8]

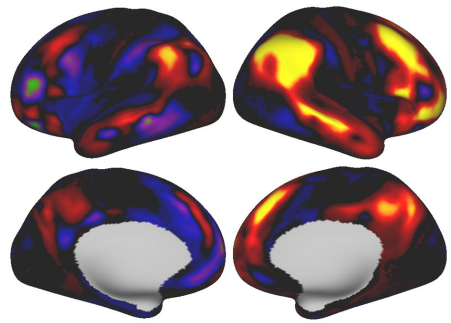
$$\boldsymbol{Y} = \boldsymbol{F}\Lambda' + E$$



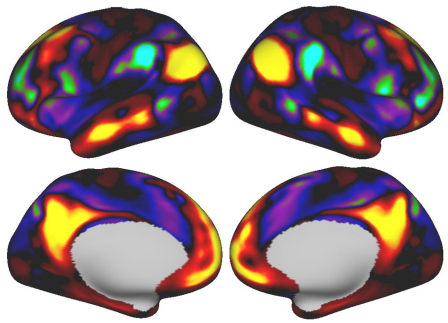
(a) PFM 13 (0.86). Dorsal attention (DA).



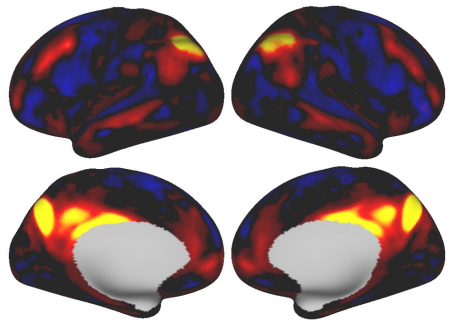
(b) PFM 18 (0.95). Left fronto-parietal control (lFPC).



(c) PFM 17 (0.77). Right fronto-parietal control (rFPC).



(d) PFM 15 (0.90). Default mode (DM).



(e) PFM 16 (0.98). Posteriomedial cortex, inferior parietal lobule (PMPL).

Posterior mean

Table of Contents

Canonical Correlation Analysis

Linear combination such that maximises the correlation

$$\underset{(P_1 \times N)}{Y_1} = \underset{(P_1 \times K)}{\Lambda_1} \times \underset{(K \times N)}{F} + \epsilon_1$$

$$\underset{(P_2 \times N)}{Y_2} = \underset{(P_2 \times K)}{\Lambda_2} \times \underset{(K \times N)}{F} + \epsilon_2$$

CCA Application: UKBB

Go to power point

Canonical Correlation Analysis (CCA)

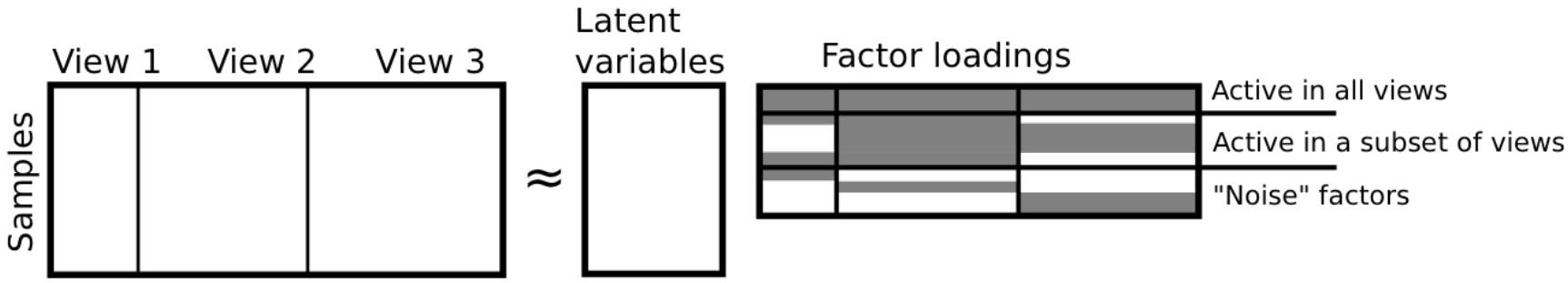
Bayesian CCA [9]

$$\begin{aligned} \boldsymbol{y}_i^{(1)} &= \boldsymbol{A}^{(1)}\boldsymbol{f}_i^{(0)} + \boldsymbol{B}^{(1)}\boldsymbol{f}_i^{(1)} + \boldsymbol{\epsilon}_i^{(1)}, \\ \boldsymbol{y}_i^{(2)} &= \boldsymbol{A}^{(2)}\boldsymbol{f}_i^{(0)} + \boldsymbol{B}^{(1)}\boldsymbol{f}_i^{(2)} + \boldsymbol{\epsilon}_i^{(2)}, \end{aligned}$$

where $\boldsymbol{f}_i^{(0)} \in \mathbb{R}^{k_0 \times 1}$ is the shared latent variable; $\boldsymbol{f}_i^{(1)} \in \mathbb{R}^{k_1 \times 1}$ and $\boldsymbol{f}_i^{(2)} \in \mathbb{R}^{k_2 \times 1}$ are the view specific latent variables.

- $\boldsymbol{\epsilon}_i^{w} \sim \mathcal{N}(\boldsymbol{0}, \Psi^{(w)})$, $w = 1, 2$
- $\boldsymbol{f}_i^{(w)} \sim \mathcal{N}(\boldsymbol{0}, I)$
- Marginal distributions: $\mathrm{cov}(\boldsymbol{y}_i)^{(w)} = \Sigma^{(w)} = \boldsymbol{B}^{(w)}\boldsymbol{B}'^{(w)} + \Psi^{(w)}$

Multi-view Learning



Generative model: Extend CCA to more than 2 views [10]

$$
\begin{aligned}
\boldsymbol{y}_i^{(w)} &= \boldsymbol{A}^{(w)}\boldsymbol{f}_i^{(0)} + \boldsymbol{B}^{(w)}\boldsymbol{f}_i^{(w)} + \boldsymbol{\epsilon}_i^{(w)}, \quad \text{for} \quad w = 1, \cdots, m \\
\boldsymbol{y}_i &= \Lambda \boldsymbol{f}_i + \boldsymbol{\epsilon}_i, \\
\Lambda &= \begin{bmatrix} \boldsymbol{A}^{(1)} & \boldsymbol{B}^{(1)} & \cdots & \boldsymbol{0} \\ \boldsymbol{A}^{(2)} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \boldsymbol{A}^{(m)} & 0 & 0 & \boldsymbol{B}^{(m)} \end{bmatrix}
\end{aligned}
$$

Multi-view Learning[10]

- Column-wise shrinkage: ARD prior on the loading columns

$$\begin{aligned} \lambda_{jh}^{w} &\sim \mathcal{N}\left(0, \left(\alpha_{h}^{(w)}\right)\right) \\ \alpha_{h}^{(w)} &\sim Ga(a_0, b_0) \end{aligned}$$

- Caveats:
  - Structured covariance matrix: Large m intractable
  - Can not capture covariance among arbitrary subset of data
  - ARD prior: Only induces column-wise sparsity

Multi-view Learning

$$\boldsymbol{y}_i = \Lambda \boldsymbol{f}_i + \boldsymbol{\epsilon}_i,$$

Structured sparsity prior on loadings [11]: sparse and dense factors

$$
\begin{aligned}
\text{Global} \quad \{ \quad & \gamma^{(w)} \sim Ga(f, \nu) \\
& \eta^{(w)} \sim Ga(e, \gamma^{(w)}) \\
\text{Factor} - \text{specific} \quad \{ \quad & \tau_h^{(w)} \sim Ga(d, \eta^{(w)}) \\
& \phi_h^{(w)} \sim Ga(c, \tau_h^{(w)}) \\
\text{Local} \quad \{ \quad & \delta_{jh}^{w} \sim Ga(b, \phi_h^{(w)}) \\
& \theta_{jh}^{(w)} \sim \pi^{(w)} Ga(a, \delta_{jh}^{w}) + (1 - \pi^{(w)}) \delta_{\phi_h^{(w)}}(.) \\
\lambda_{jh}^{w} \quad \sim \quad & \mathcal{N}(0, \theta_{jh}^{(w)})
\end{aligned}
$$

Bibliography I

[1] Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[2] T. W. Anderson and Herman Rubin. Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry*, pages 111–150, Berkeley, Calif., 1956. University of California Press.

[3] Joshua Chan, Roberto Leon-Gonzalez, and Rodney W. Strachan. Invariant inference and efficient computation in the static factor model. *Journal of the American Statistical Association*, 0(ja):0–0, 2017.

Bibliography II

[4] Hedibert Freitas Lopes and Mike West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14(1):41–67, 2004.

[5] Sylvia Kaufmann and Christian Schumacher. Bayesian estimation of sparse dynamic factor models with order-independent identification. Technical Report 13.04, April 2013.

[6] David Knowles and Zoubin Ghahramani. Nonparametric bayesian sparse factor models with application to gene expression modeling. *Ann. Appl. Stat.*, 5(2B):1534–1552, 06 2011. doi: 10.1214/10-AOAS435.

[7] Veronika RoÄ■kovÃą and Edward I. George. Fast bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622, 2016.

Bibliography III

[8] Samuel J. Harrison, Mark W. Woolrich, Emma C. Robinson, Matthew F. Glasser, Christian F. Beckmann, Mark Jenkinson, and Stephen M. Smith. Large-scale probabilistic functional modes from resting state fmri. *NeuroImage*, 109:217 – 231, 2015.

[9] Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14(Apr):965–1003, 2013.

[10] Arto Klami, Seppo Virtanen, Eemeli Leppäaho, and Samuel Kaski. Group factor analysis. *IEEE transactions on neural networks and learning systems*, 26(9):2136–2147, 2015.

[11] Shiwen Zhao, Chuan Gao, Sayan Mukherjee, and Barbara E Engelhardt. Bayesian group factor analysis with structured sparsity. *The Journal of Machine Learning Research*, 17(1): 6868–6914, 2016.