Transfer Learning via Bayesian Latent Factor Analysis

Preliminary Exam Presentation

Liz Lorenzi[1]
Joint work with Katherine Heller[1] and Ricardo Henao[2]

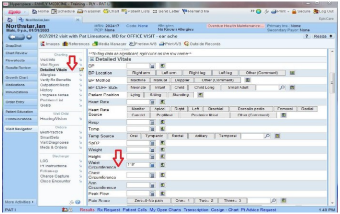
Duke University Department of Statistical Science[1],
Duke University Department of Electrical and Computer Engineering[2]

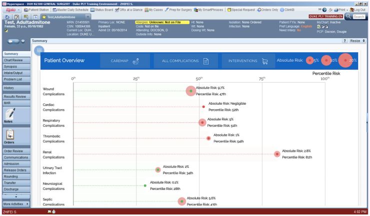
November 30, 2016

kēlaHealth

The Problem:

- Average surgical complication rate is 15%
- 50% of these complications are avoidable
- Average cost of complication is $11,626 [Dimick et al., 2004]





Our mission:

- Predict post-operative complications using surgery patient electronic health records (EHRs)
- Enhance decision making of clinicians by suggesting appropriate interventions

Leveraging information across databases

- National Surgery Quality Improvement Program (NSQIP)
  - 3.7 million patients, > 700 hospitals contribute
- Duke Medical Center
  - 13,711 patients

Programs collect same information but have different populations

- Duke:
  - teaching hospital
  - higher variability in outcomes and complications
  - more experimental surgeries
- NSQIP:
  - wide variety in hospital types
  - different patient care and patient cohorts

Our goal:

1. Predict complications for patients at Duke
2. Leverage information in NSQIP
3. Discern important factors in predicting complications

Transfer Learning

In machine learning, we define a problem with an additional source of information *(NSQIP)* apart from the standard training data *(Duke)* to be **transfer learning** [Pan and Yang, 2010].

- **Goal**: improve learning in target task by leveraging knowledge from related tasks
- **Our approach**: Hierarchical latent factor models
  - Learn one set of latent factors that accounts for the distributional differences across populations
  - Appropriately model separate covariance structure for each population

Latent Factor Model (LFM)

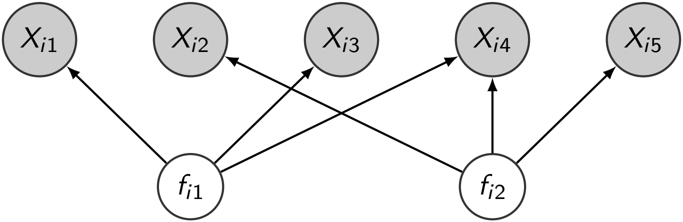
LFM explains underlying variability among observed, correlated covariates in lower-dimensional unobserved "factors."

- Relate observed data, X_i, to a k-vector of random variables, f_i

$$X_i = \Lambda f_i + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \Sigma)$$

$$\Sigma = \mathrm{diag}(\sigma_1^2, ...\sigma_p^2)$$

- Λ is $P \times K$ factor loadings matrix



Properties of LFM

We assume the P variables of X are distributed as a multivariate zero-mean normal distribution

$$X \sim \text{Normal}(0, \Omega)$$

where $\Omega = \Lambda\Lambda' + \Sigma$

- Dependence between observed variables is induced by marginalizing over the distribution of the factors
- Allows for direct modeling of covariance matrix

Transfer Learning via Latent Factor Model (TL-LFM)

- $\{X_i^t : i = 1, ..., n_t\}$ represent predictors of target data
- $\{X_i^s : i = 1, ..., n_s\}$ represent predictors of source data

$$X_i^j = \Lambda^j f_i + \epsilon_i$$

- where $j \in \{s, t\}$ represents the different populations

We facilitate sharing between groups via the prior setup:

$$m_p \sim \mathrm{N}(0, \frac{1}{\phi} I_k)$$

$$\Lambda_p^s \sim \mathrm{N}(m_p, \frac{1}{\phi_s} I_K), \quad \Lambda_p^t \sim \mathrm{N}(m_p, \frac{1}{\phi_t} I_K)$$

Properties of TL-LFM

Marginalizing over the factors, X has the following form:

$$X_i \sim \mathrm{N}_p(0, \Omega^j)$$

$$\Omega^j = V(X_i | \Lambda^j, \Sigma) = \Lambda^j \Lambda^{j\prime} + \Sigma$$

- Results in separate modeling of populations' covariances

TL-LFM Regression

Let $Z = \{Y, X\}$ represent the full data.

- Joint model implies that $E(y_i|x_i) = x_i'\theta^j$ where $\theta^j = {\Omega^j_{XX}}^{-1}\Omega^j_{YX}$

The posterior predictive distribution is easily found by solving,

$$f(y_{n+1}|y_1, ..., y_n, x_{n+1}) = \int f(y_{n+1}|x_{n+1}, \Omega)\pi(\Omega|y_1, ..., y_n, x_1, ..., x_n)d\Omega$$

Simulation Experiments

Goal: mimic transfer learning across two populations

- different sample ratios (target:source)
- 35 binary predictors, 35 continuous predictors
- repeated ten times

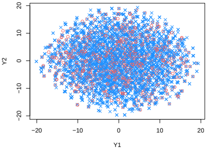
Simulate Z_i, for $i = 1, ..., 5000$ from a 70-dimensional normal distribution, with zero mean and covariance equal to Ω^j.

For each population:

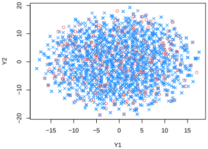
- Sample each row of Λ^j from a Normal$(0, I_k)$ with $K = 20$
- Randomly select two locations of first row of Λ and set to -1 and 1, with the rest 0
- Draw the diagonal of Σ from an InvGamma(1, 0.5) with prior mean equal to 2.

Visualizing TL-LFM

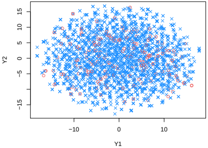
Plot K-dimensional latent factors using t-sne (van der Maaten, Hinton 2008) comparing hierarchical and non-hierarchical models



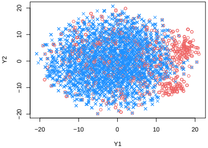
(a) TL - 700:2800



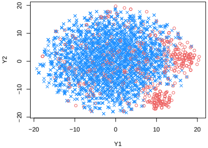
(b) TL - 500:2500



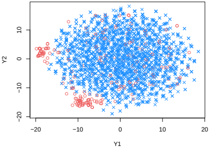
(c) TL - 100:2000



(d) NoTL - 700:2800



(e) NoTL - 500:2500



(f) NoTL - 100:2000

Figure: Method - Target:Source

Evaluating TL-LFM Prediction Results

We report the area under the ROC curve with standard errors

Target:Source	TL-LFM	LFM	Lasso
700:2800	**0.809** (.007)	0.587 (.005)	0.723 (.006)
500:2500	**0.790** (.005)	0.594 (.008)	0.732 (.005)
200:2000	**0.744** (.005)	0.547 (.005)	0.585 (.004)

Table: Tested on target only held out test set

Surgery Data Results

NSQIP/Duke data contains information for a single patient undergoing surgery, with covariates describing

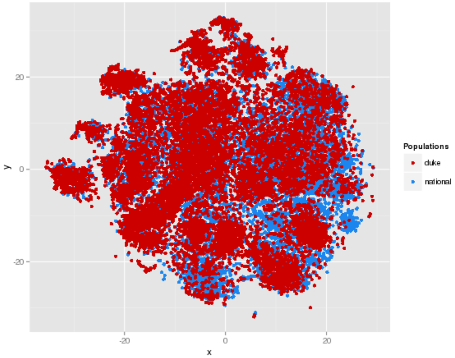
- demographic information
- preoperative and intraoperative variables
- outcomes of surgery (e.g. cardiac arrest, pneumonia, infection)

Results show Lasso outperforms our TL-LFM.

TL-LFM	LFM	Lasso
0.73	0.60	**0.76**

Table: Prediction on Duke-only patients for any-morbidity

Results of TL-LFM: How to improve performance?



We focus on 3 areas:

1. Modeling modal structure
2. Allowing more flexible transferring of information
3. Inducing stronger sparsity

1-2: Hierarchical Dirichlet Process

$$G^0 | H \sim \mathrm{DP}(\alpha_0, H)$$

Base measure of child DP is also DP.

$$G^j | G^0 \sim \mathrm{DP}(\alpha_j, G^0), \quad \forall j \in \{1, .., J\}$$

$$G^0 = \sum_{k=1}^{\infty} \pi_k^0 \delta_{\lambda_{pk}^0}$$

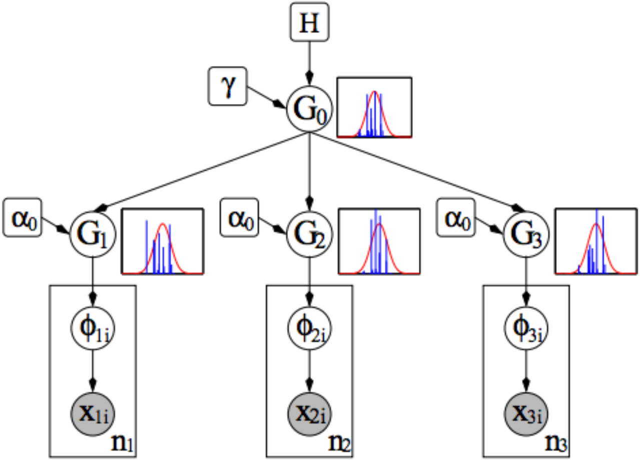
$$\lambda_p^0 \sim \mathrm{Norm}(0, \Sigma_\lambda)$$

and for each group $j \in J$,

$$G^j = \sum_{k=1}^{\infty} \pi_k^j \delta_{\lambda_{pk}^0}$$

[Teh et al., 2012]

Graphical model for HDP



[Teh et al., 2012]

Finite HDP Conversion

HDP is infinite limit of finite mixture models formulation.

$$\pi^0 | \alpha_0 \sim \mathrm{Dir}(\alpha_0 / K, ..., \alpha_0 / K)$$

$$\pi^j | \alpha_j, \pi^0 \sim \mathrm{Dir}(\alpha_j \pi^0)$$

$$\lambda^0 \sim \mathrm{Normal}(0, \Sigma_\lambda)$$

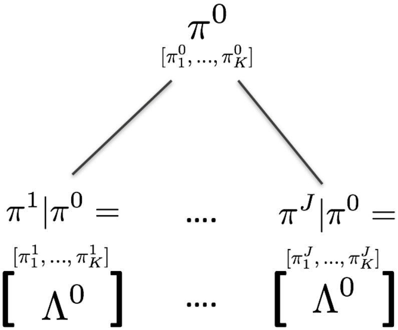
$$G^j = \sum_{k=1}^{K} \pi_k^j \delta_{\lambda_k^0}$$

Adapting HDP for factor model

Instead of drawing from discrete mixture:

$$G^j = \sum_{k=1}^{K} \pi_k^j \delta_{\lambda_k^0}$$

Consider the $P \times K$ loadings matrix for λ^0 weighted by the stick-breaking weights, $\Lambda^j = [\pi_1^j \lambda_1^0, .., \pi_K^j \lambda_K^0]$.



HDP as scale mixture

We use the stick-breaking proportions from the HDP as a weighting scheme to the rows of the loadings matrix.

$$\sqrt{\pi^j}\lambda_p^0$$

where $\lambda_p^0 \sim \mathrm{N}(0, \frac{1}{\phi} \cdot I_K)$.
This results in

$$\sqrt{\pi_k^j}\lambda_k^0 \sim \mathrm{N}(0, \pi_k^j \frac{1}{\phi})$$

Can we formulate this as a sparse prior to address our third goal?

3: Sparse modeling

From Bayesian-learning perspective, there are 2 main sparse-estimation options

- Discrete mixtures - e.g. spike and slab ([Mitchell and Beauchamp, 1988]; [George and McCulloch, 1993])

$$\beta_j \sim w \cdot g(\beta_j) + (1 - w) \cdot \delta_0$$

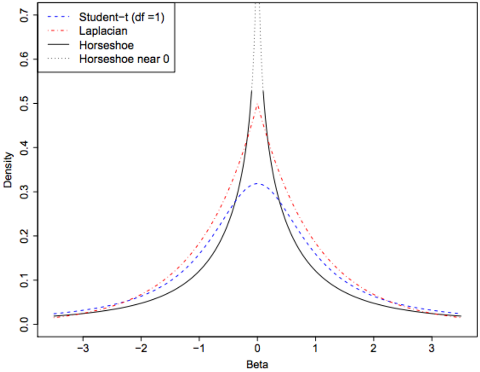
- Shrinkage priors - e.g. horseshoe, L1/Laplace prior ([Carvalho et al., 2009]; [Tibshirani, 1996]; [Mohamed et al., 2011])

$$\beta_j | \tau^2, \lambda_j \sim \text{Norm}(0, \tau^2 \lambda_j^2)$$

$$\lambda_j^2 \sim \pi(\lambda_j^2)$$

$$\tau^2 \sim \pi(\tau^2)$$

Examples of Scale Mixture Priors



Marginal distributions for β:

- Student-t with $\lambda_j \sim \text{InvGam}(v/2, v/2)$
- Double Exponential/Laplace with $\lambda_j \sim \text{Exp}(2)$
- Horseshoe with half Cauchy, $\lambda_j \sim \text{C}^+(0, 1)$

[Carvalho et al., 2009]

How to choose a sparse prior?

[Polson and Scott, 2010] presents criteria for evaluating different sparsity priors. They focus on two guidelines:

- $\pi(\lambda_j^2)$ should have heavy tails
- $\pi(\tau^2)$ should have substantial mass at zero

"Strong global shrinkage handles the noise; the local λ_j's act to detect the signals."

Motivating new model: TL-SLFM

Back to the original model. Let j represent separate populations (expanding from just S or T):

$$\mathbf{X}_{ji} = \Lambda^j f_{ji} + \epsilon_{ji}$$

$$\epsilon_{ji} \sim \mathrm{N}(0, \Sigma_j), \quad \Sigma_j = \mathrm{diag}(\sigma^2_{j1}, ..., \sigma^2_{jP})$$

$$f_{ji} \sim \mathrm{N}(0, I_K)$$

How can we change the prior on Λ^j to result in covariance structure that adjusts to our goals?

Constructing Λ

For each k in $1, .., K$, we weigh a global λ_k^0 with $\sqrt{\pi_k^j}$, such that $\lambda_k^j | \pi_k^j := \sqrt{\pi_k^j} \lambda_k^0$.

- The global parameter, π^j, controls shrinkage of λ_p.

$$\lambda_p^j | \pi^j \sim \text{Normal}(0, \frac{1}{\phi_p} \pi^j I_k)$$

$$\pi^j | \pi^0 \sim \text{Dir}(\alpha_j \pi^0)$$

$$\pi^0 \sim \text{Dir}(\alpha_0 / K)$$

- The local parameter, ϕ_p, will have heavy tails. For $\phi_{pk} \in \phi_p$,

$$\phi_{pk} \sim \text{Gamma}(\tau/2, \tau/2)$$

Properties of resulting factor model

Model learns a marginal covariance of $\Omega^j = \lambda^{j\prime}\lambda^j + \Sigma^j$, where λ^j is the resulting sparse loadings matrix.

Results in partitioned covariance that adjusts to each population:

$$\Omega^j = (\lambda^0 \Pi^j \lambda^{0\prime}) + \Sigma^j$$

where $\Pi^j = \mathrm{diag}(\pi_1^j, ..., \pi_K^j)$

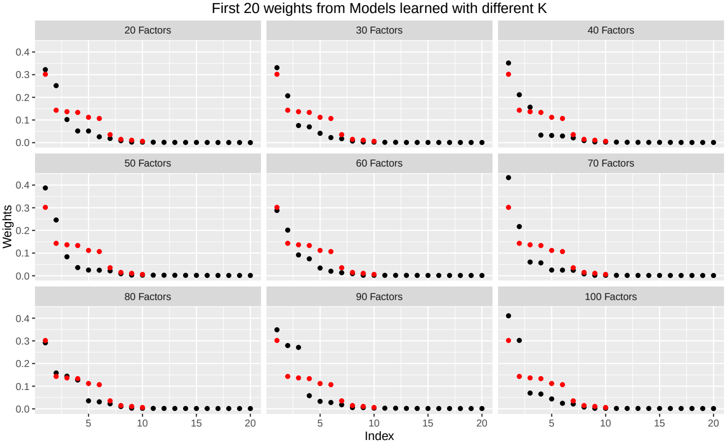
Choosing number of factors

Choosing correct number of factors is difficult computationally and conceptually.

- Early work chooses number of factors by maximizing marginal likelihood, AIC, or BIC
- [Lopes and West, 2004] suggest a reversible-jump MCMC method to learn K
- [Lucas et al., 2006]; [Carvalho et al., 2012] choose number of factors by using model selection priors to zero out parts of the loadings matrix
- [Bhattacharya and Dunson, 2011] propose a multiplicative gamma shrinkage prior to allow the number of factors to approach infinity while the columns of the loadings matrix increasingly shrink towards zero

Model is robust to choosing number of factors

Comparing data simulated under a 10-factor model (red) to 20 largest weights learned from models with $K = 20 : 100$ (black).



Models appropriately shrink weights for models with $K > 10$.

Deriving Inference for stick-breaking scale mixture

We use the following identity to decompose the weights, π^j:

$$w_k^j \sim \text{Gam}(\alpha_j \pi^0, 1), \quad \pi^j = \left(\frac{w_1^j}{\sum w_k^j}, ..., \frac{w_K^j}{\sum w_k^j}\right) \sim \text{Dir}(\alpha_j \pi_1^0, .., \alpha_j \pi_K^0)$$

We rewrite the generative model using the unnormalized w^j.

$$X_{ji} = \lambda^j f_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \Sigma_j), \quad \Sigma_j = diag(\sigma_{jp}^2)_{p \in \{1,...,P\}}$$
$$\sigma_{jp}^2 \sim InvGam(\nu/2, \nu s^2/2)$$
$$f_i \sim N(0, I_K)$$

$$\lambda_p^j | w^j \sim Normal(0, W^j 1/\phi_p), \quad W^j = \text{diag}(w_1^j, .., w_K^j)$$
$$w_k^j | \alpha_j, \pi^0 \sim Gamma(\alpha_j \pi_k^0, 1)$$
$$\pi^0 \sim Dirichlet(\alpha_0/K)$$
$$\phi_{pk} \sim Gamma(\tau/2, \tau/2)$$

where $i = 1, .., n,\ p = 1, .., P,\ k = 1, .., K,\ j = 1, .., J.$

Resulting Full Conditionals for stick-breaking scale mixture

Results in the following tractable full conditionals:

$$(\lambda_p^j|-) \sim \mathrm{N}(m = (\sigma_{jp}^{-2} F' X_{jp}) V, V = (\phi_p W^{j-1} + \sigma_{jp}^{-2} F' F)^{-1})$$

$$(w_k^j|-) \sim \mathrm{GIG}(p = \alpha_j \pi_k^0 - P/2, a = 2, b = \Phi_k(\lambda_k^{jT} \lambda_k^j))$$

$$(\phi_{pk}|-) \sim \mathrm{Gamma}(\tau/2 + J/2, \tau/2 + \sum_{j=1}^{J} \frac{\lambda_{pk}^{j2}}{2 w_k^j})$$

Note: Capital parameters represent diagonal matrix

Inference: Learning π_0

For drawing π_0 we use a Metropolis-Hastings sampling scheme:
We propose π^*_{0k}:

$$\pi^*_{0k} \sim \mathrm{LogNormal}(\log(\pi_k^{t-1}), C)$$

and normalize. Then accept according to the acceptance ratio:

$$A(\pi_0^*|\pi_0^{t-1}) = \min\left(1, \frac{P(\pi_0^*|w_1, w_2)}{P(\pi_0^{t-1}|w_1, w_2)} \frac{g(\pi_0^{t-1}|\pi_0^*)}{g(\pi_0^*|\pi_0^{t-1})}\right)$$

Results in better mixing.

Initial Results: Simulation

We set up TL-SLFM as regression model.

- Reporting area under ROC curve with standard errors from 10 simulations

	TL-SLFM	TL-LFM	LFM	Lasso
700:2800	**0.812** (0.008)	0.788 (0.010)	0.754 (0.012)	0.783 (0.009)
500:2500	**0.791** (0.010)	0.765 (0.012)	0.694 (0.008)	0.762 (0.006)
200:2000	**0.795** (0.008)	0.744 (0.010)	0.668 (0.011)	0.698 (0.010)

Table: Prediction on target-only held out set.

Initial Results: Real Data

Until inference is scaled to evaluate the full data, we test on subsets of the data by surgery.

- Hernia surgeries (5000 in NSQIP to 362 in Duke)

TL-SLFM	TL-LFM	Lasso
0.876	0.733	0.838

Table: Prediction on Duke-only patients for any-morbidity

- Breast Mastectomy (5000 in NSQIP to 680 in Duke)

TL-SLFM	TL-LFM	Lasso
0.747	0.698	0.706

Table: Prediction on Duke-only patients for any-morbidity

Final words:

Overview/Takeaways

- Presented a transfer learning framework using latent factor models
- Extended framework for more complicated relationships between populations through TL-SLFM
- Created a novel way to use stick-breaking weights in a scale mixture

Next steps

Transfer Learning

- Scale inference method using stochastic variational Bayes or Stochastic Gradient Descent MCMC
- Extend SLFM to be nonparametric (infinite number of factors)
- Apply to different problems for multiple populations with varying types of information

Next steps

Causal Inference

- kelaHealth
  - Measure effectiveness of kelaHealth in reducing complications
  - Consider more tuned intervention based on expected individual treatment effect
- MS Mosaic
  - Learn sequential treatment effect for MS patients for varying types of treatments

Thank you!

Prelim Committee:

- Katherine Heller, Ph.D.
- Ricardo Henao, Ph.D.
- Fan Li, Ph.D.
- Surya Tokdar, Ph.D

kelaHealth Team:

- Bora Chang, M.D. Candidate
- Erich Huang, M.D./Ph.D.
- Ouwen Huang, M.D./Ph.D. Candidate
- Jeff Sun, M.D.

Works Cited: I

Bhattacharya, A. and Dunson, D. B. (2011).
Sparse bayesian infinite factor models.
Biometrika, 98(2):291–306.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2012).
High-dimensional sparse factor modeling: Applications in gene expression genomics.
Journal of the American Statistical Association.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009).
Handling sparsity via the horseshoe.
In *AISTATS*, volume 5, pages 73–80.

Dimick, J. B., Chen, S. L., Taheri, P. A., Henderson, W. G., Khuri, S. F., and Campbell, D. A. (2004).
Hospital costs associated with surgical complications: a report from the private-sector national surgical quality improvement program.
Journal of the American College of Surgeons, 199(4):531–537.

George, E. I. and McCulloch, R. E. (1993).
Variable selection via gibbs sampling.
Journal of the American Statistical Association, 88(423):881–889.

Works Cited: II

Lopes, H. F. and West, M. (2004).
Bayesian model assessment in factor analysis.
Statistica Sinica, pages 41–67.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2006).
Sparse statistical modelling in gene expression genomics.
Bayesian Inference for Gene Expression and Proteomics, 1:0–1.

Mitchell, T. J. and Beauchamp, J. J. (1988).
Bayesian variable selection in linear regression.
Journal of the American Statistical Association, 83(404):1023–1032.

Mohamed, S., Heller, K., and Ghahramani, Z. (2011).
Bayesian and l1 approaches to sparse unsupervised learning.
arXiv preprint arXiv:1106.1157.

Pan, S. J. and Yang, Q. (2010).
A survey on transfer learning.
IEEE Transactions on knowledge and data engineering, 22(10):1345–1359.

Polson, N. G. and Scott, J. G. (2010).
Shrink globally, act locally: Sparse bayesian regularization and prediction.
Bayesian Statistics, 9:501–538.

Works Cited: III

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2012).
Hierarchical dirichlet processes.
Journal of the american statistical association.

Tibshirani, R. (1996).
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288.

Appendix: More on Stick-breaking prior

Alternatively can write prior as product of two random variables:

$$\Lambda_k^j := \sqrt{w_k^j}\lambda_k^0$$

- Results in marginal covariance of $\Omega^j = (\lambda^{0\prime} W^j \lambda^0) + \Sigma^j$
- This product results in the following distribution for the element Λ_{hk}^j where $h = 1, ..., P$.

$$f(\Lambda_{hk}^j) = f(\sqrt{w_k^j}\lambda_k^0) = \frac{\phi^{-1/2+\alpha\pi_k}}{2^{1/2-\alpha\pi_k-1}\pi^{1/2}\Gamma(\alpha\pi_k)}(\Lambda_{hk})^{3\alpha\pi_k}\mathcal{K}_{\alpha\pi_k}(\sqrt{1/\phi 2\Lambda_{hk}^2})$$

- where $\mathcal{K}$ is a modified Bessel function of the second kind.
- Connection: This product distribution is very similar to the marginal distribution of $f(\beta_k)$ from the generalized double pareto scale mixture (Caron, Doucet, 2008).