Dimension Reduction with Heavy Tails

Gabriel Kuhn
Munich University of Technology
http/www.ma.tum.de/stat

4th Conference on Extreme Value Analysis
Gothenburg, August 15.–19., 2005

Reference: Klüppelberg, C. and Kuhn, G. (2005) Dimension Reduction with Heavy Tails. *In preparation*.

Factor Model

- Observable d-dimensional random vector $\boldsymbol{X}$
- *Model:* $\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{f} + \mathbf{V}\boldsymbol{e}$
  – $\mathbf{L}\boldsymbol{f}$: k-dimensional non-observable *common factors* $\boldsymbol{f}$, loading matrix $\mathbf{L}$
  – $\mathbf{V}\boldsymbol{e}$: *specific factors* $\boldsymbol{e}$, diagonal matrix $\mathbf{V}$
  – $(\boldsymbol{f}, \boldsymbol{e})$ are uncorrelated (independent)
  Idea: Distribution of $\boldsymbol{X}$ described by linear combination of k factors with componentwise extra source of randomness.
- *Classical model:* $(\boldsymbol{f}, \boldsymbol{e}) \sim \mathcal{N}_{k+d}(\mathbf{0}, \mathbf{I})$ and $\mathrm{Cov}(\boldsymbol{X}) =: \boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{V}^2$
  Disadvantages: – Data may not be normal
  – No heavy-tailed model
  – Dependence in extremes cannot be modeled
  – Margins of the same type
- **Task:** Overcome disadvantages above

Distribution Models

- *Elliptical Distribution:* $\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + G\mathbf{A}\boldsymbol{U}^{(k)}$

  $G > 0$ independent of $\boldsymbol{U}^{(k)} \sim \mathrm{unif}\{\boldsymbol{s} : \|\boldsymbol{s}\| = 1\}$, $\mathbf{A} \in \mathbb{R}^{d\times k}$, $\boldsymbol{\Sigma} := \mathbf{A}\mathbf{A}^T$

  $E(\boldsymbol{X}) = \boldsymbol{\mu}$, $\mathrm{Cov}(\boldsymbol{X}) = EG^2\boldsymbol{\Sigma}/k$, $\mathrm{Corr}(\boldsymbol{X}) = \mathrm{diag}(\boldsymbol{\Sigma})^{-1/2}\boldsymbol{\Sigma}\mathrm{diag}(\boldsymbol{\Sigma})^{-1/2} =: \mathbf{R}$

- **Example:** – *normal:* $\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \sqrt{\chi^2_k}\mathbf{A}\boldsymbol{U}^{(k)} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

  – *multivariate t_ν:* $\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \sqrt{\nu\chi^2_k/\chi^2_\nu}\mathbf{A}\boldsymbol{U}^{(k)} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \sqrt{\nu/\chi^2_\nu}\mathcal{N}_d(\mathbf{0}, \boldsymbol{\Sigma})$

Extended Factor Model

Drop normal assumption and consider (with $\mathbf{L}\mathbf{L}^T + \mathbf{V}^2 = \mathbf{\Sigma}$):

$\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \mathbf{L}\boldsymbol{f} + \mathbf{V}\boldsymbol{e}$ is elliptical: $\mathbf{L}\boldsymbol{f} + \mathbf{V}\boldsymbol{e} = (\mathbf{L}, \mathbf{V}) \begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{e} \end{pmatrix} \stackrel{\mathrm{d}}{=} G(\mathbf{L}, \mathbf{V})\boldsymbol{U}^{(k+d)}$,

e.g. choose multivariate t_ν-distribution

Remark:
- $\boldsymbol{f}$ and $\boldsymbol{e}$ are uncorrelated but not independent
- alternatively: same dependence structure, but arbitrary margins (copula approach)
- $P(G > x) \sim Cx^{-\nu}$ needed for modelling tail dependence

Factorization

- Standard approach uses (normal) ml algorithm for decomposition $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{V}^2$
  Definition: $f_{\mathrm{ml}}^{(k)}(\mathbf{A}) = (\mathbf{L}, \mathbf{V}), \quad (f_{\mathrm{ml}}^{(k)}(\mathbf{A}))^2 := (\mathbf{L}, \mathbf{V})(\mathbf{L}, \mathbf{V})^T = \mathbf{L}\mathbf{L}^T + \mathbf{V}^2$

- Lemma: $\mathbf{\Sigma}_n \xrightarrow{P} \mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{V}^2$ and neighborhood of $\mathbf{\Sigma}$ decomposable by $f_{\mathrm{ml}}^{(k)}$
  $\Rightarrow \left(f_{\mathrm{ml}}^{(k)}(\mathbf{\Sigma}_n)\right)^2 \xrightarrow{P} \mathbf{\Sigma}$

- Interpretation: Given some *consistent* and *composable* covariance (correlation) estimator, the algorithm computes a consistent decomposition (independent of distribution model)

- Remark: In application algorithm almost always produces a decomposition

Dependence Concepts

- *Kendall's τ:* Let $(X, Y) \overset{iid}{\sim} (\tilde{X}, \tilde{Y})$

$$\tau := P\Big((X - \tilde{X})(Y - \tilde{Y}) > 0\Big) - P\Big((X - \tilde{X})(Y - \tilde{Y}) < 0\Big)$$

  Elliptical distribution $\Rightarrow \mathbf{R} = \sin(\pi \mathbf{T}/2), \quad \mathbf{T} = (\tau_{ij})_{1 \le i,j \le d}$

- *Tail Dependence:* (X, Y) with margins F_X, F_Y

$$\lambda := \lim_{u \searrow 0} P\left(Y < F_Y^{\leftarrow}(u) \,|X < F_X^{\leftarrow}(u)\right)$$

  Elliptical distribution and $P(G > x) \sim Cx^{-\nu}$, $\nu \in (0, \infty)$

$$\Rightarrow \mathbf{R} = 1 - 2\frac{\left(F_{t,\nu+1}^{\leftarrow}(1 - \Lambda/2)\right)^2}{\nu + 1 + \left(F_{t,\nu+1}^{\leftarrow}(1 - \Lambda/2)\right)^2}$$

  $\mathbf{\Lambda} = (\lambda_{ij})_{1 \le i,j \le d}$ and $F_{t,\nu}$: df of 1-dim t_ν

- Remark: Both dependence concepts independent of margins

Example

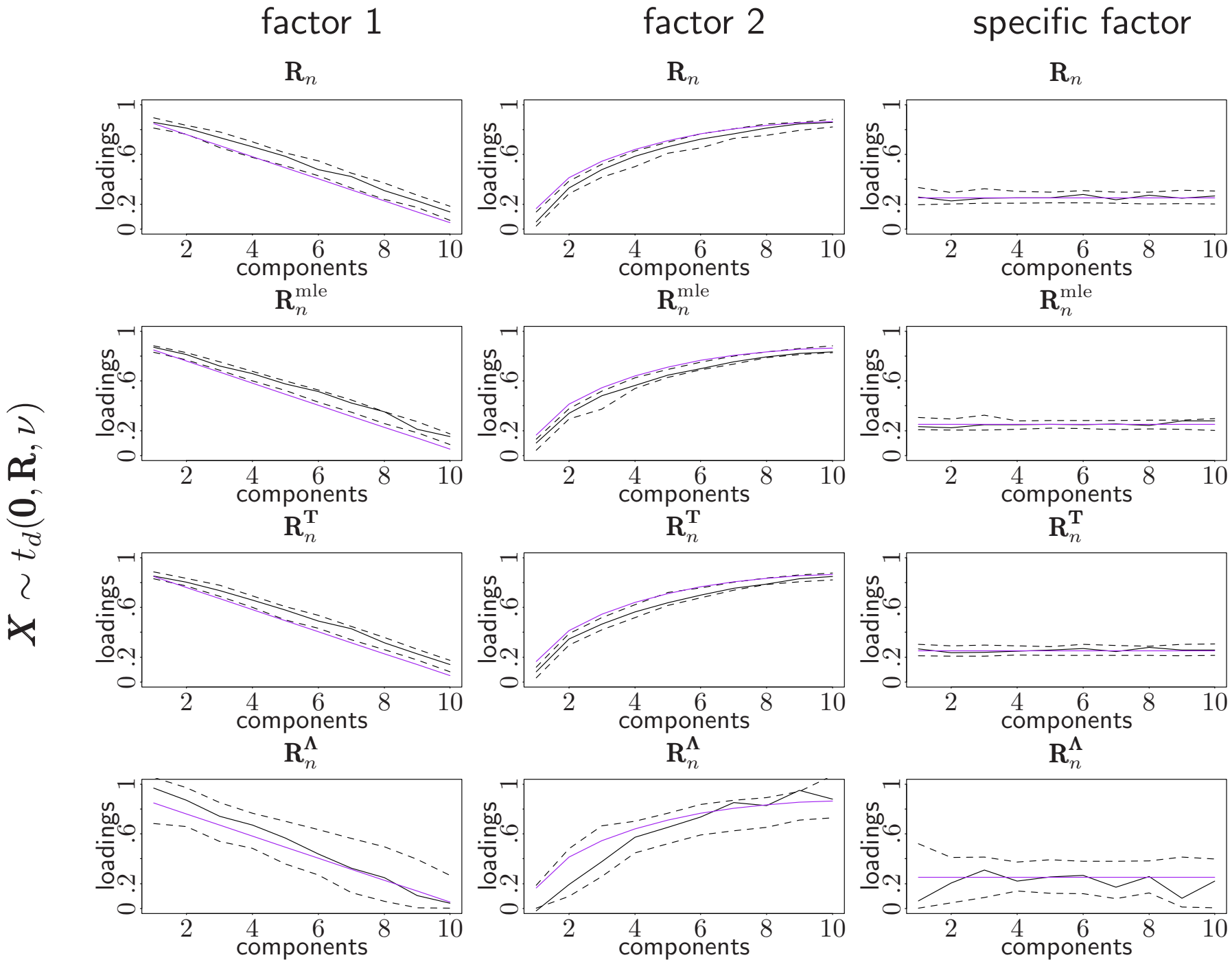
- Choose $d = 10$, $k = 2$ factors with loadings

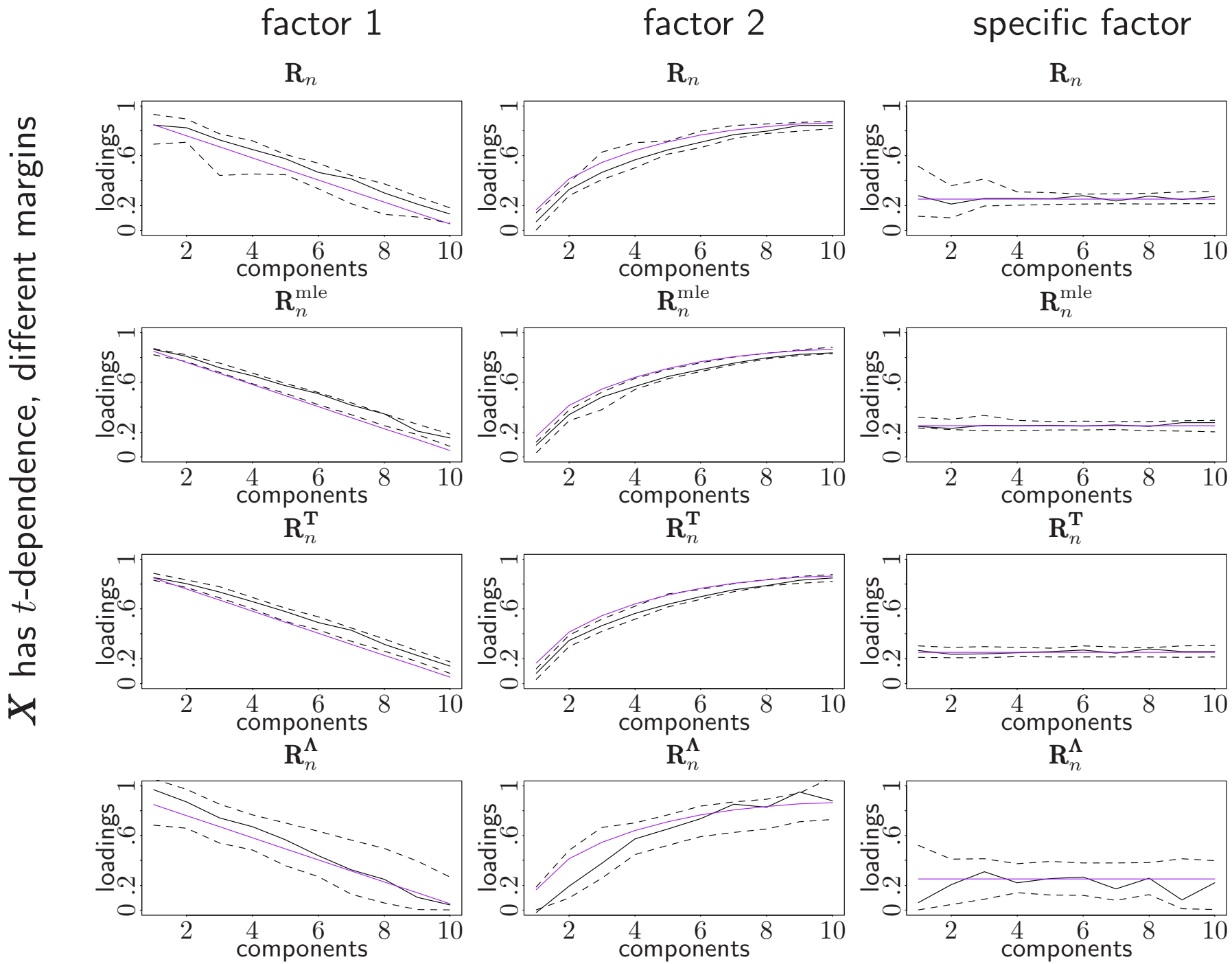
component	1	2	3	4	5	6	7	8	9	10
$\mathbf{L}_{\cdot,1}$	.85	.76	.67	.58	.49	.41	.32	.23	.14	.05
$\mathbf{L}_{\cdot,2}$	.17	.41	.55	.64	.71	.77	.81	.84	.85	.86
$\mathrm{diag}(\mathbf{V}^2)$	.25	.25	.25	.25	.25	.25	.25	.25	.25	.25

  ($\Rightarrow \mathbf{L}\mathbf{L}^T + \mathbf{V}^2 = \mathbf{R}$ is a correlation matrix)

- consider factor model(s) (given before)
  1. $\boldsymbol{X} \stackrel{\mathrm{d}}{=} \mathbf{L}\boldsymbol{f} + \mathbf{V}\mathbf{e} \sim t_d(\mathbf{0}, \mathbf{R}, \nu)$ with $\nu = 6$
  2. $\boldsymbol{X}$ has same $t(\mathbf{0}, \mathbf{R}, \nu)$ dependence structure, but different margins $F_i = t_{\nu_i}$, $\boldsymbol{\nu} = (3, \ldots, 10)$

- Simulation length $n = 2000$, repeat 500 times
  Plots of different estimation methods and 95%-CI's of loadings


factor 1
factor 2
specific factor
$\mathbf{R}_n$
$\mathbf{R}_n^{\text{mle}}$
$\mathbf{R}_n^{\mathbf{T}}$
$\mathbf{R}_n^{\mathbf{\Lambda}}$
loadings
components
$\boldsymbol{X} \sim t_d(\mathbf{0}, \mathbf{R}, \nu)$

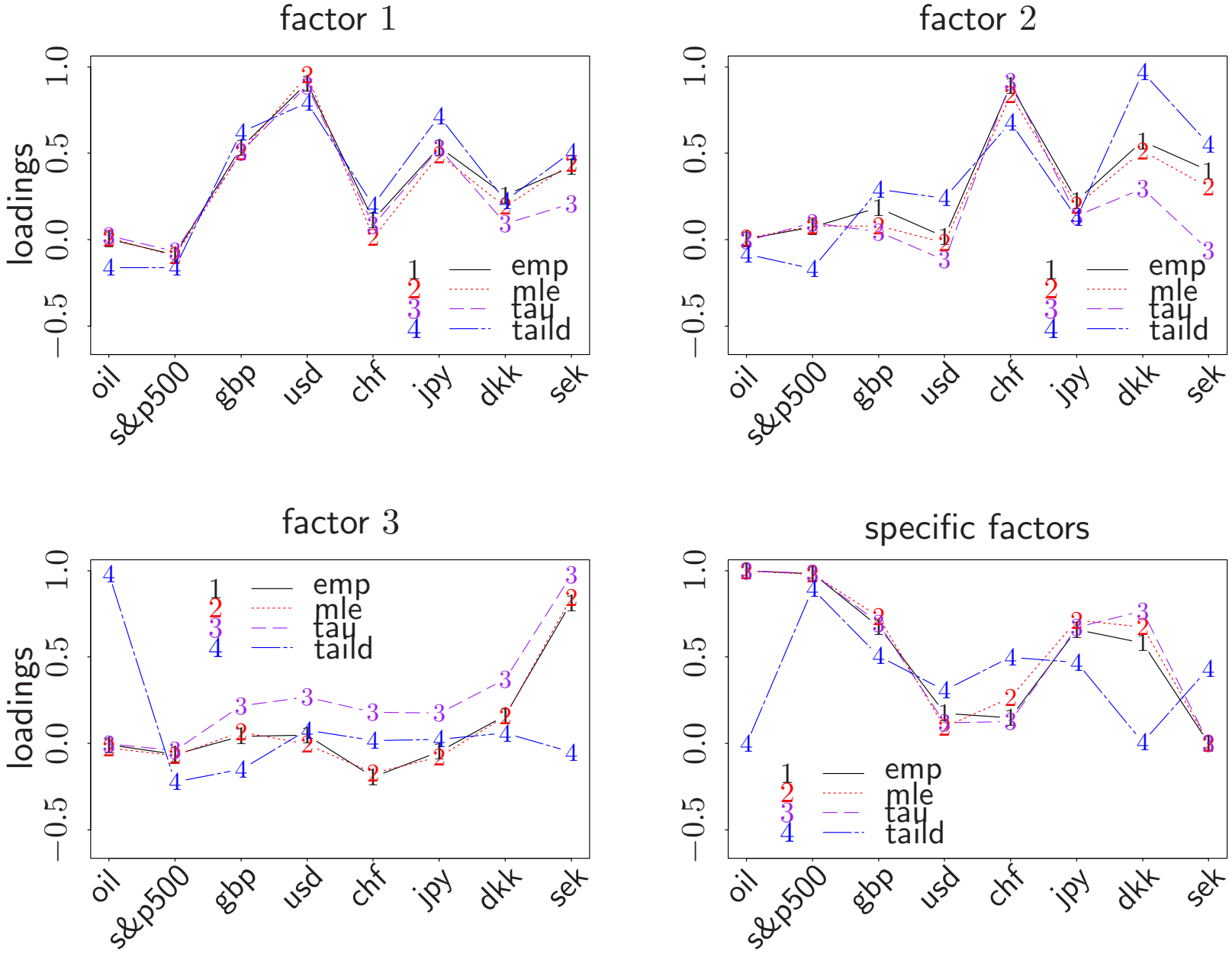

factor 1
factor 2
specific factor
$\boldsymbol{X}$ has t-dependence, different margins
$\mathbf{R}_n$
$\mathbf{R}_n^{\mathrm{mle}}$
$\mathbf{R}_n^{\mathbf{T}}$
$\mathbf{R}_n^{\mathbf{\Lambda}}$
loadings
components
0
.2
.6
1
2
4
6
8
10

Example

- Consider 8-dimensional set of data:
  oil, s&p500, gbp, usd, chf, jpy, dkk and sek (exchange rates w.r.t. euro)

- Daily log-returns between May, 1985 to June, 2004 (n=4904)

- Apply factor analysis with different estimators as before


factor 1
factor 2
factor 3
specific factors
loadings
−0.5
0.0
0.5
1.0
oil
s&p500
gbp
usd
chf
jpy
dkk
sek
emp
mle
tau
taild