Fitting Multidimensional Latent Variable Models using an Efficient Laplace Approximation

Dimitris Rizopoulos
Department of Biostatistics, Erasmus University Medical Center, the Netherlands

d.rizopoulos@erasmusmc.nl

Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

January 13th, 2010

1.1 Introduction

- Item Response Theory (IRT) plays nowadays a central role in the analysis and study of tests and item scores

- Application of IRT models can be found in many fields
  - ▷ psychometrics
  - ▷ educational sciences
  - ▷ sociometrics
  - ▷ medicine
  - ▷ ...

1.1 Introduction (cont'd)

- Standard IRT models are available in special-purpose software, such as BILOG & MULTILOG and in R

- For R more information can be found at: http://cran.r-project.org/web/views/Psychometrics.html

1.1 Introduction (cont'd)

- A fundamental assumption behind these standard IRT models is *unidimensionality*:
  - ▷ the interdependencies between the responses of each sample unit are explained by a *single* latent variable

- In some cases tests are designed to measure a single trait, e.g.,
  - ▷ reading ability
  - ▷ environmental attitude
  - ▷ . . .

1.1 Introduction (cont'd)

- However, in many cases unidimensionality is too strict to be true, e.g.,
  - ▷ tests measure different latent traits
    - * mathematics test: algebra, calculus, etc.
    - * types of depression: major depressive disorder, dysthymia, manic depression
  - ▷ hierarchical/multilevel designs
    - * subjects are nested within clusters
    - * items are nested within different dimensions

1.1 Introduction (cont'd)

- If there is a predominant general factor in the data, and dimensions beyond that major dimension are relatively small, then multidimensionality has a little effect on derived inferences

- However, if the unidimensionality assumption is seriously violated, then
  - ▷ item parameter estimates will be biased, and
  - ▷ the standard errors associated with ability parameter estimates will be too small

1.2 Motivating Case Study

- Programme for International Student Assessment (PISA)
  - ▷ launched by the Organization for Economic Co-operation and Development
  - ▷ collect data on student and institutional factors that can explain differences in student performance
  - ▷ in 2003, 41 countries participated and the survey covered mathematics, reading, science, and problem solving

- Data features
  - ▷ different dimensions: ability in mathematics, reading, science, problem solving
  - ▷ hierarchical design: students nested in schools, schools nested in countries

1.2 Motivating Case Study

- Aim: estimate item and ability parameters, taking into account covariates and the hierarchical design

- Using a multilevel analysis we will be able to simultaneously estimate the item and ability

1.3 What is this talk about

- Problem: as we will illustrate fitting complex latent variable models is a computationally challenging task requiring a lot of computing time

- Our Aim: develop a computationally flexible approach that can fit latent variable models with complex latent structures in reasonable computing time

- Work in progress. . . (no results yet available)
  - ▷ promising results from the relevant framework of joint models for longitudinal and time-to-event data (with high-dimensional random effects)

2 Multidimensional IRT Models

- Notation:
  - ▷ $\boldsymbol{y}_i$: vector of responses for the ith subject
  - ▷ $\boldsymbol{z}_i$: vector of latent variables

- Basic assumption: conditional independence (CI)
  - ▷ given the latent structure, we assume that the responses of the ith subject are independent

$$p(\boldsymbol{y}_i \mid \boldsymbol{z}_i) = \prod_{k=1}^{p} p(y_{ik} \mid \boldsymbol{z}_i)$$

where $p(\cdot)$ denotes a pdf

2 Multidimensional IRT Models (cont'd)

- In order CI to hold, a complex latent structure may be required

- A general definition of an IRT model

$$g\{E(\boldsymbol{y}_i \mid \boldsymbol{z}_i)\} = \boldsymbol{X}_i \boldsymbol{\beta}^{(x)} + \boldsymbol{Z}_i \boldsymbol{\beta}^{(z)}$$

where

- ▷ $g(\cdot)$: link function
- ▷ $\boldsymbol{X}_i$: design matrix for covariates
- ▷ $\boldsymbol{Z}_i$: vector of latent variables
- ▷ $\boldsymbol{\beta}^{(x)}$: regression coefficients for covariates
- ▷ $\boldsymbol{\beta}^{(z)}$: regression coefficients for latent variables

2 Multidimensional IRT Models (cont'd)

- Examples:
  - ▷ dichotomous data – 1 level (i subject, k item)

$$\text{logit}\{\Pr(y_{ik} = 1 \mid \boldsymbol{z}_i, \boldsymbol{\theta})\} = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_q z_{iq}$$

q-latent-variable model

2 Multidimensional IRT Models (cont'd)

- Examples:
  - ▷ polytomous data (c = 1, 2, . . .) – 2 levels (i subject in group j, k item)

$$\text{Pr}(y_{ijk} = c \mid \boldsymbol{z}_i, \boldsymbol{\theta}) = \text{expit}(a_k z_{ij} - b_{k,c-1}) - \text{expit}(a_k z_{ij} - b_{k,c})$$

$$\text{Level I:} \quad z_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + e_{ij}$$

$$\text{Level II:} \quad \begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} w_{1j} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} w_{1j} + u_{1j} \end{aligned}$$

e_{ij}, $\boldsymbol{u}_i$ denote Error Terms

3.1 ML Estimation

- Estimation of multidimensional IRT model is typically based on marginal maximum likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int p(\boldsymbol{y}_i \mid \boldsymbol{z}_i; \boldsymbol{\theta}) \, p(\boldsymbol{z}_i; \boldsymbol{\theta}) \, d\boldsymbol{z}_i$$

  where

  ▷ $\boldsymbol{\theta}$ denotes the parameter vector

  ▷ $p(\boldsymbol{y}_i \mid \boldsymbol{z}_i; \boldsymbol{\theta})$ denoted the density of the multidimensional IRT as introduced above

  ▷ we assume that $\boldsymbol{z}_i$ are distributed according to a parametric distribution

  ▷ we integrate $\boldsymbol{z}_i$ to obtain the marginal distribution for the observed responses

3.1 ML Estimation (cont'd)

- Due to the fact that the integral

$$\int p(\boldsymbol{y}_i|\boldsymbol{z}_i)\, p(\boldsymbol{z}_i)\, d\boldsymbol{z}_i$$

  does not have a closed form solution

- Maximization of $\ell(\boldsymbol{\theta})$ is a computationally challenging task – it requires a combination of
  - ▷ numerical integration, and
  - ▷ numerical optimization

3.1 ML Estimation (cont'd)

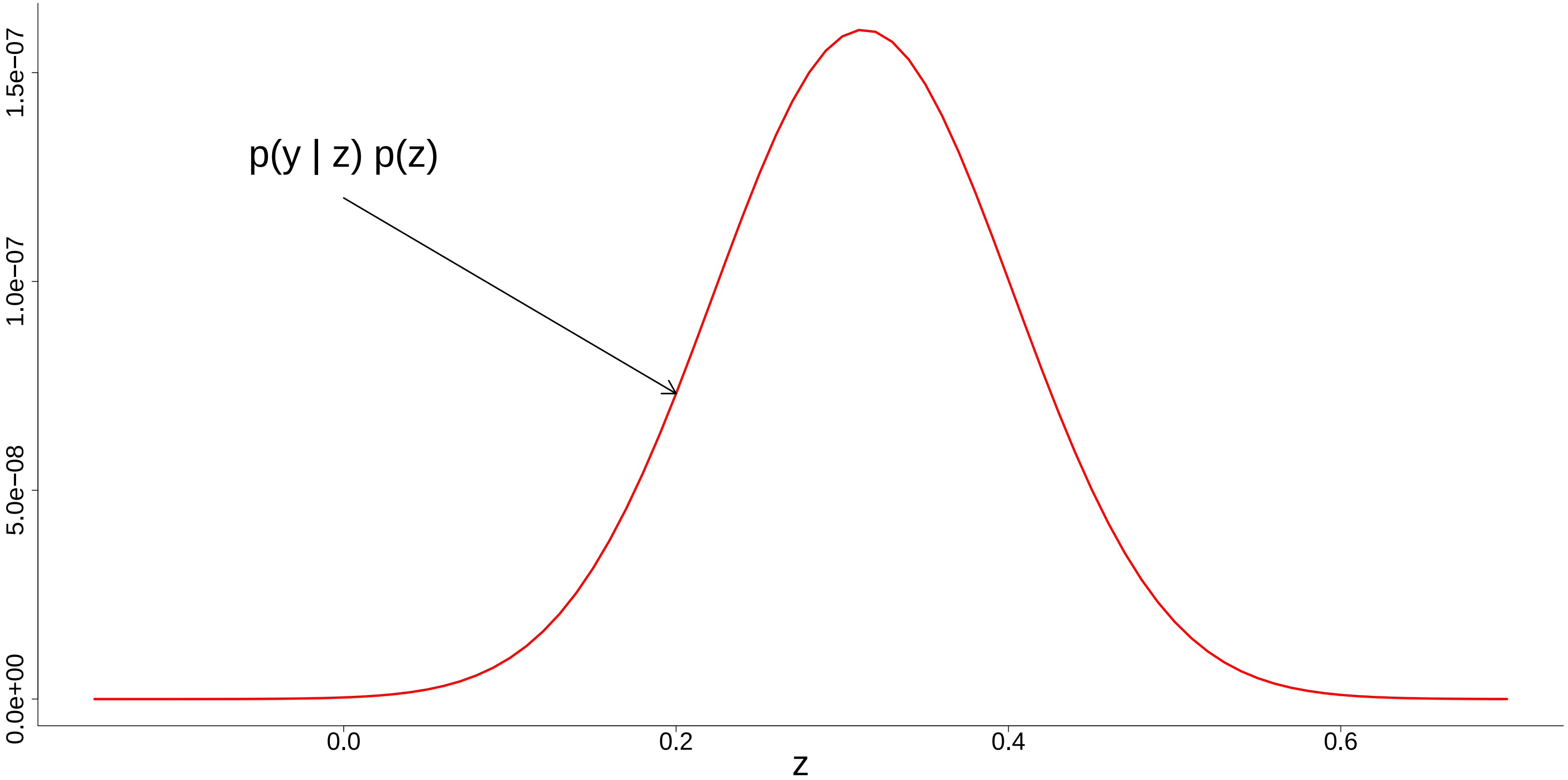
- For numerical optimization standard choices are
  - ▷ EM algorithm (we treat $\boldsymbol{z}_i$ as 'missing values')
  - ▷ Newton-type algorithms, such as Newton-Raphson or quasi-Newton

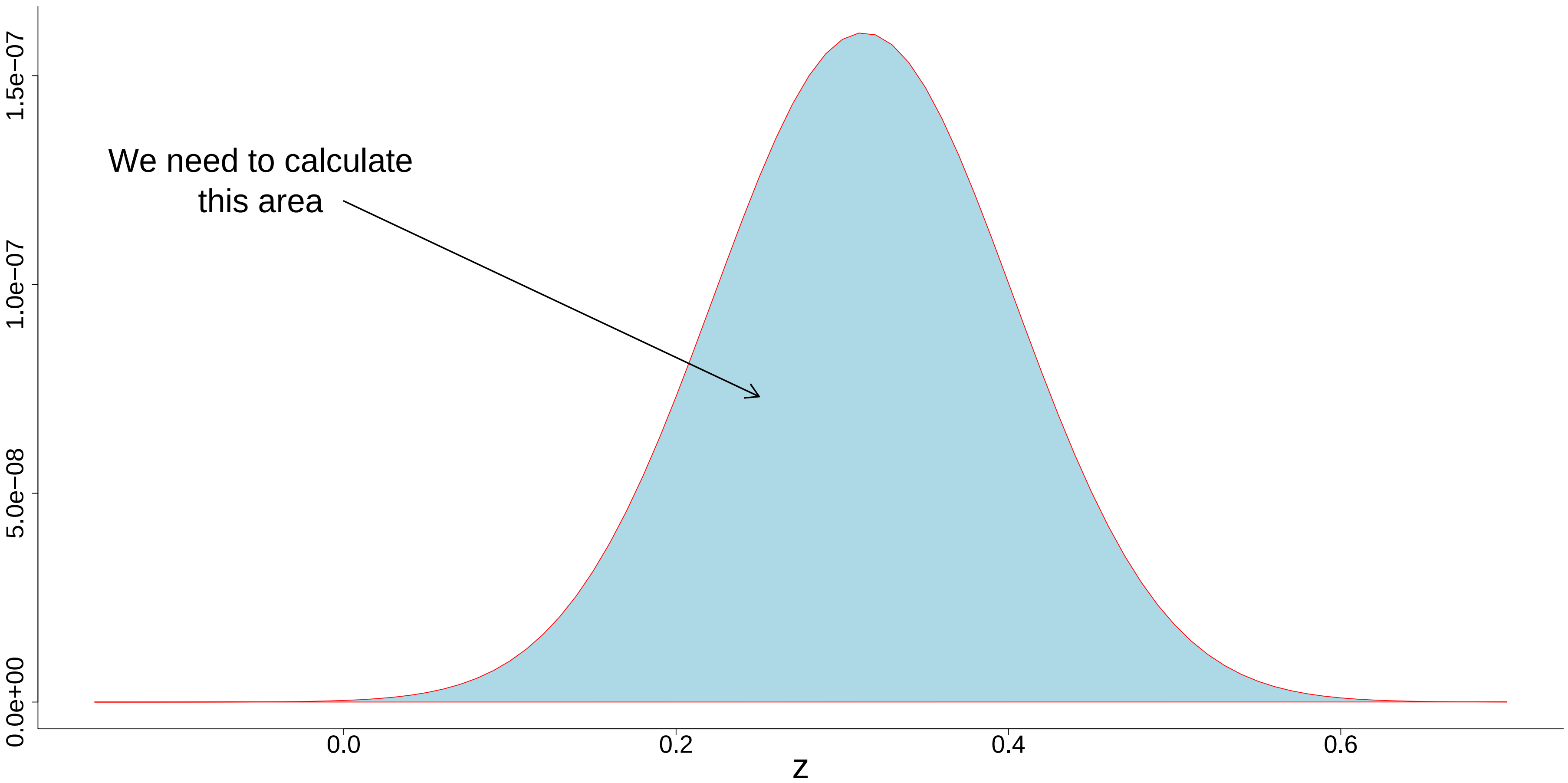
- Hybrid approaches that start with EM (as a refinement of the starting values) for a fixed number of iterations, and continue with quasi-Newton have also been successfully used

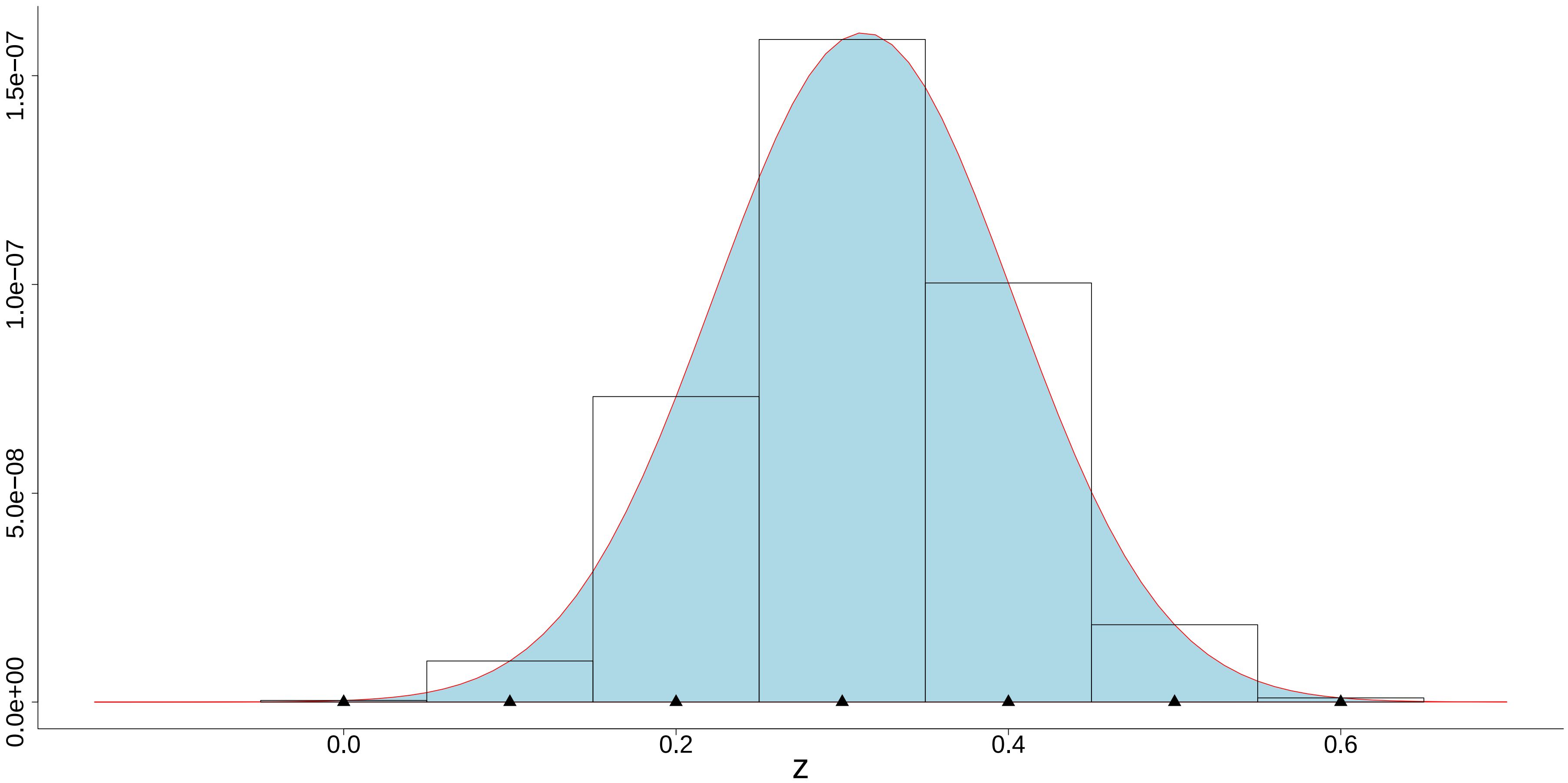
3.1 ML Estimation (cont'd)

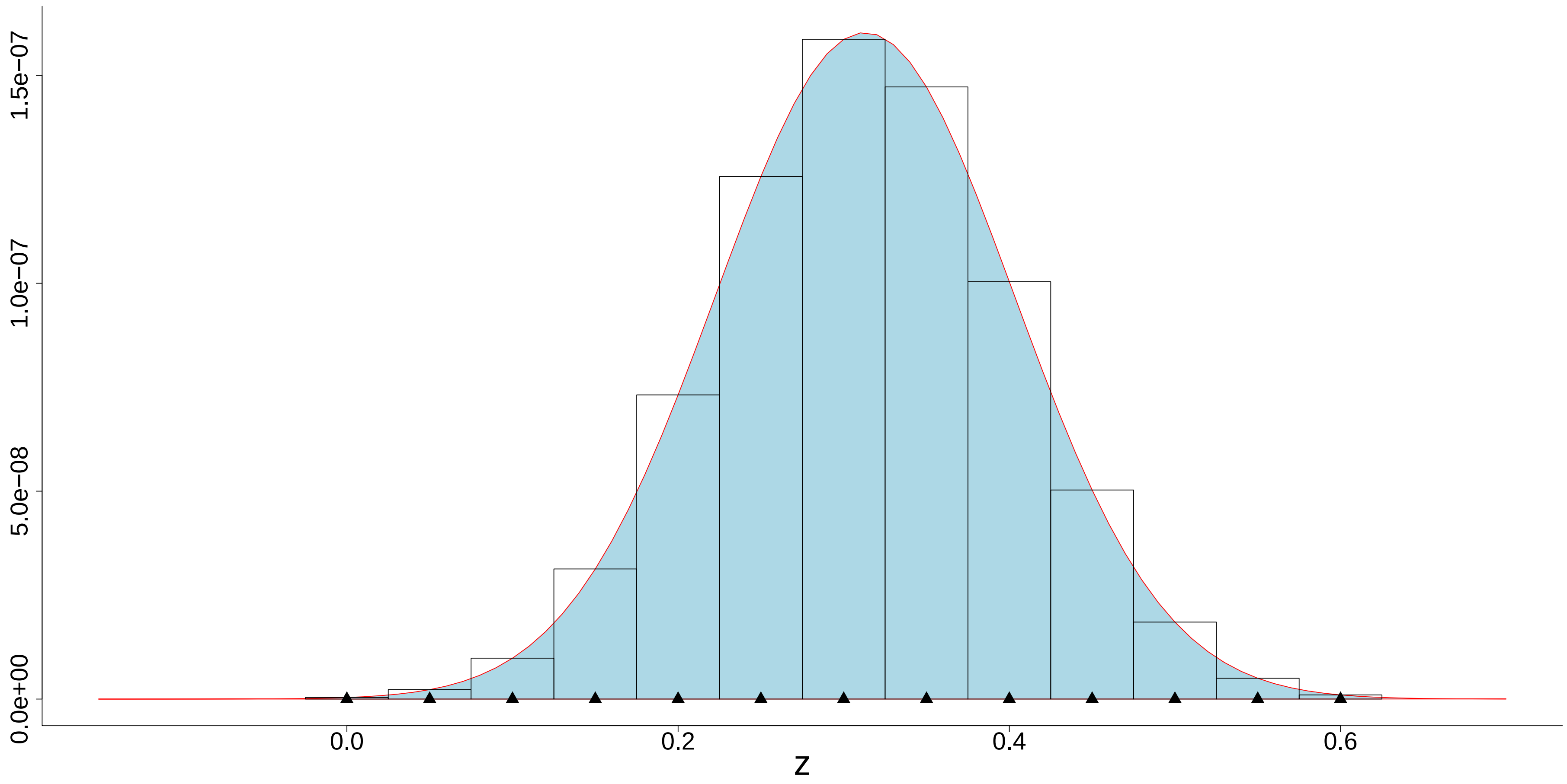
- For numerical integration standard choices are
  - ▷ Monte Carlo
  - ▷ (adaptive) Gauss-Hermite quadrature rule

- However, these are prohibitive when a moderate to high number of latent variables is considered

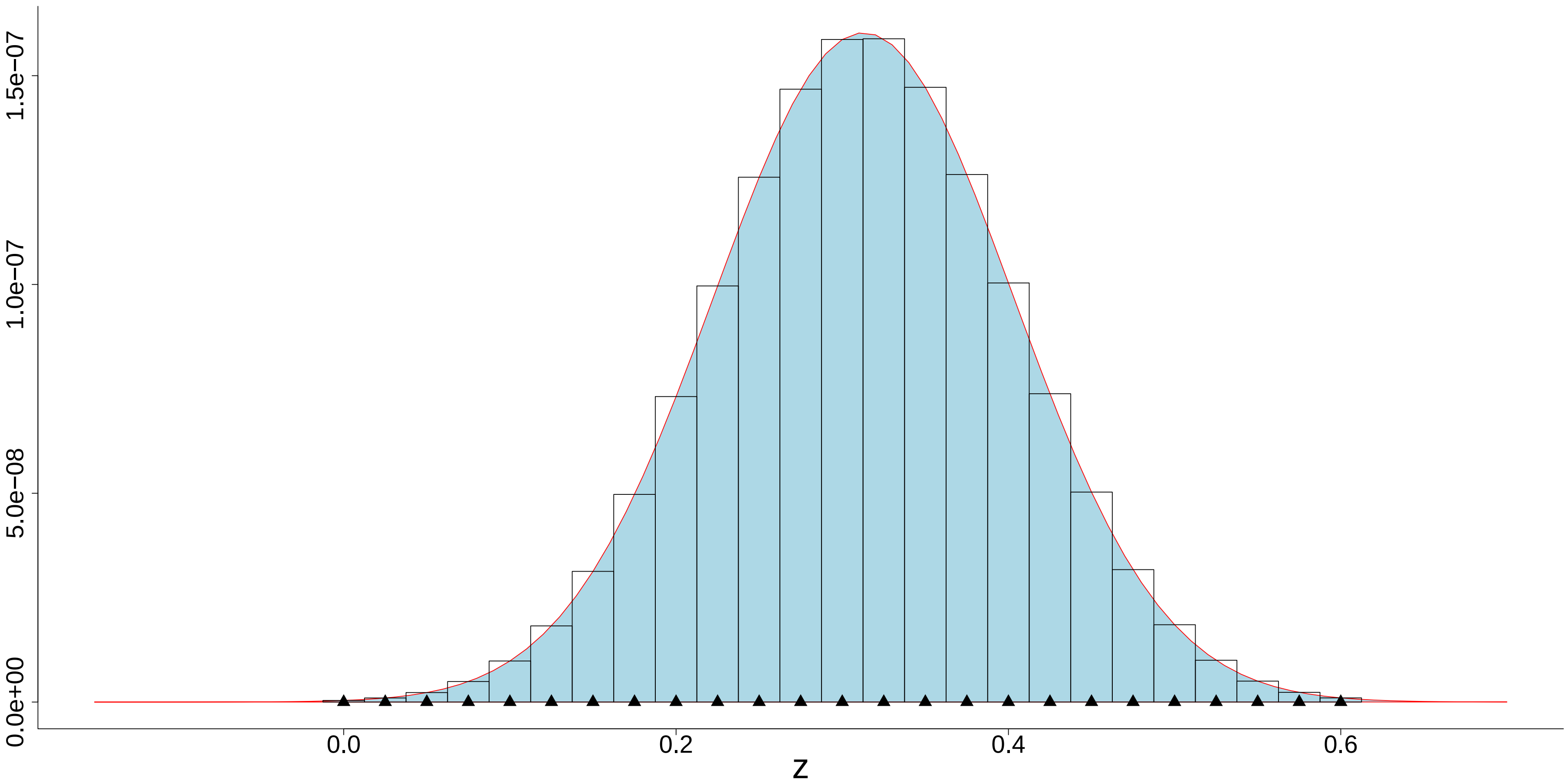

p(y | z) p(z)
1.5e−07
1.0e−07
5.0e−08
0.0e+00
0.0
0.2
0.4
0.6
z


We need to calculate
this area
1.5e−07
1.0e−07
5.0e−08
0.0e+00
0.0
0.2
0.4
0.6
z


1.5e−07
1.0e−07
5.0e−08
0.0e+00
0.0
0.2
0.4
0.6
z

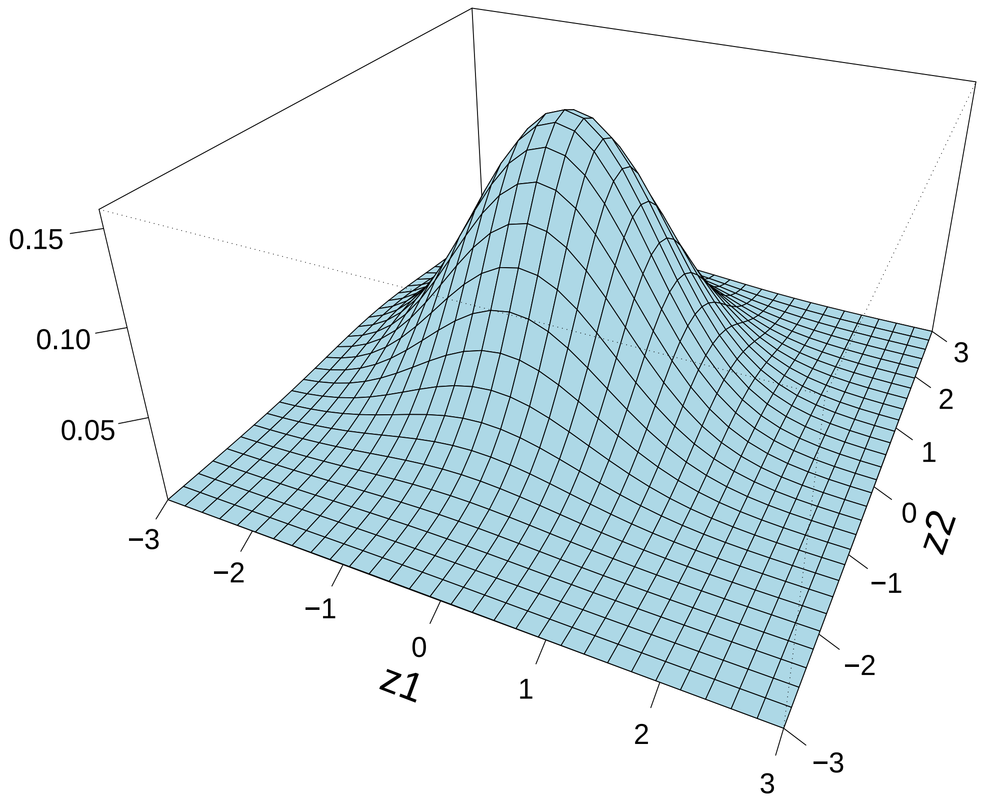


1.5e−07
1.0e−07
5.0e−08
0.0e+00
0.0
0.2
0.4
0.6
z



1.5e−07
1.0e−07
5.0e−08
0.0e+00
0.0
0.2
0.4
0.6
z

Two Latent Variables



3.2 Laplace Approximation

- An alternative solution instead of numerical integration is the Laplace approximation

$$p(\boldsymbol{y}_i; \boldsymbol{\theta}) = \int \exp\{\log p(\boldsymbol{y}_i \mid \boldsymbol{z}_i; \boldsymbol{\theta}) + \log p(\boldsymbol{z}_i; \boldsymbol{\theta})\}\, d\boldsymbol{z}_i$$

$$= \Big[(2\pi)^{q/2}\mathsf{det}(\boldsymbol{\Sigma})^{-1/2} \exp\{\log p(\boldsymbol{y}_i \mid \widehat{\boldsymbol{z}}_i; \boldsymbol{\theta}) + \log p(\widehat{\boldsymbol{z}}_i; \boldsymbol{\theta})\}\Big] \big(1 + O(p_i^{-1})\big),$$

where

- ▷ $\widehat{\boldsymbol{z}}_i = \underset{\boldsymbol{z}_i}{\operatorname{argmax}}\{\log p(\boldsymbol{y}_i \mid \boldsymbol{z}_i) + \log p(\boldsymbol{z}_i)\}$
- ▷ $\boldsymbol{\Sigma} = -\nabla^2 \{\log p(\boldsymbol{y}_i \mid \boldsymbol{z}_i) + \log p(\boldsymbol{z}_i)\}\Big|_{\boldsymbol{z}_i = \widehat{\boldsymbol{z}}_i}$

3.2 Laplace Approximation (cont'd)

- It requires a large number of repeated measurements per individual in order to provide a good approximation to the integral

- Contrary to Monte Carlo and Gaussian quadrature, in the Laplace approximation we cannot control the approximation error

- Therefore, it would be desirable to improve the approximation, especially for small to moderate number of repeated measurements per individual

3.3 Score Vector in Latent Variable Models

- The score vector in latent variable models can be written in the form (Rizopoulos et al., JRSSB, 2009)

$$\begin{aligned} S_i(\boldsymbol{\theta}) &= \sum_i \frac{\partial}{\partial \boldsymbol{\theta}} \log \int p(\boldsymbol{y}_i \mid \boldsymbol{z}_i; \boldsymbol{\theta})\, p(\boldsymbol{z}_i; \boldsymbol{\theta})\, d\boldsymbol{z}_i \\ &= \sum_i \int \frac{\partial}{\partial \boldsymbol{\theta}} \Big\{ \log p(\boldsymbol{y}_i \mid \boldsymbol{z}_i; \boldsymbol{\theta}) + \log p(\boldsymbol{z}_i; \boldsymbol{\theta}) \Big\}\, p(\boldsymbol{z}_i \mid \boldsymbol{y}_i; \boldsymbol{\theta})\, d\boldsymbol{z}_i \end{aligned}$$

- Observed data score vector = expected value of complete data score vector *wrt* the posterior of the latent variables given the observed data

3.3 Score Vector in Latent Variable Models (cont'd)

- Why is this useful
  - ▷ easy to combine EM with quasi-Newton
  - ▷ enables a more efficient Laplace approximation

3.4 EM & quasi-Newton

- EM algorithm for latent variable models
  - ▷ maximize the expected value of the complete data log-likelihood (expectation is taken *wrt* the posterior of the latent variables given the observed data)

$$Q_i(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*) \; = \; \int \log\{p(\boldsymbol{y}_i \mid \boldsymbol{z}_i; \boldsymbol{\theta}) p(\boldsymbol{z}_i; \boldsymbol{\theta})\} \; p(\boldsymbol{z}_i \mid \boldsymbol{y}_i; \boldsymbol{\theta}^*) \; d\boldsymbol{z}_i$$

- To maximize $Q(\cdot)$ we need to solve

$$\int \frac{\partial}{\partial \boldsymbol{\theta}} \Big\{ \log p(\boldsymbol{y}_i \mid \boldsymbol{z}_i; \boldsymbol{\theta}) + \log p(\boldsymbol{z}_i; \boldsymbol{\theta}) \Big\} \; p(\boldsymbol{z}_i \mid \boldsymbol{y}_i; \boldsymbol{\theta}^*) \; d\boldsymbol{z}_i = 0$$

which is $S_i(\boldsymbol{\theta})$

3.4 EM & quasi-Newton (cont'd)

- Direct maximization for latent variable models using quasi-Newton
  - ▷ maximize the observed data log-likelihood $\Rightarrow$ solve the score equations $S_i(\boldsymbol{\theta}) = 0$

- Therefore, both EM and quasi-Newton require calculation of the same function $S_i(\boldsymbol{\theta})$
  - ▷ take into advantage of the stability of EM during the first iteration, and later change to quasi-Newton which has better convergence rate

3.5 Fully Exponential Laplace Approximation

- Fitting latent variable models under MML requires calculations of the form

$$\int A(\boldsymbol{z}_i)\, p(\boldsymbol{z}_i \mid \boldsymbol{y}_i)\, d\boldsymbol{z}_i,$$

where $A(\boldsymbol{z}_i) = \partial\{\log p(\boldsymbol{y}_i \mid \boldsymbol{z}_i; \boldsymbol{\theta}) + \log p(\boldsymbol{z}_i; \boldsymbol{\theta})\}/\partial\boldsymbol{\theta}$

- Note that the above can be written as

$$E\left\{A(\boldsymbol{z}_i)\right\} = \frac{\int A(\boldsymbol{z}_i)\, p(\boldsymbol{y}_i \mid \boldsymbol{z}_i)\, p(\boldsymbol{z}_i)\, d\boldsymbol{z}_i}{\int p(\boldsymbol{y}_i \mid \boldsymbol{z}_i)\, p(\boldsymbol{z}_i)\, d\boldsymbol{z}_i}$$

3.5 Fully Exponential Laplace Approximation

- If we apply the standard Laplace approximation in the numerator and denominator of $E\{A(\boldsymbol{z}_i)\}$, then the $O(p_i^{-1})$ terms cancel out, which leads to a $O(p_i^{-2})$ approximation

- This approximation has been used for Bayesian computations (Tierney et al., JASA, 1989)

- Caveat: it can only be applied for positive functions
  - ▷ however, $A(\boldsymbol{z}_i)$, which is the complete data score vector, is not restricted to be positive

3.5 Fully Exponential Laplace Approximation

- Write the previous equation as

$$E\left\{A(\boldsymbol{z}_i)\right\} = \frac{d}{ds}\log E[\exp\{sA(\boldsymbol{z}_i)\}]\Big|_{s=0}$$

- Then we obtain the approximation

$$E\left\{A(\boldsymbol{z}_i)\right\} = \left\{A(\widehat{\boldsymbol{z}}_i) + \frac{d}{ds}\log \mathsf{det}(\boldsymbol{\Sigma}_s)^{-1/2}\Big|_{s=0}\right\}\left(1 + O(p_i^{-2})\right),$$

where

▷ $\boldsymbol{\Sigma}_s = -\nabla^2\left\{sA(\boldsymbol{z}_i) + \log p(\boldsymbol{y}_i \mid \boldsymbol{z}_i) + \log p(\boldsymbol{z}_i)\right\}\Big|_{\boldsymbol{z}_i = \widehat{\boldsymbol{z}}_i^{(s)}}$

▷ $\widehat{\boldsymbol{z}}_i$ same as in the simple Laplace approximation

3.5 Fully Exponential Laplace Approximation

- The enhanced Laplace approximation is
  - ▷ the simple Laplace approximation,
  - ▷ and differentiation of $\{\log \mathsf{det}(\boldsymbol{\Sigma}_s)^{-1/2}\}$ *wrt* s

$$\frac{\partial}{\partial s_k} \log \mathsf{det}(\Sigma_s)^{-1/2} = -\frac{1}{2}\mathsf{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial}{\partial s_k}\boldsymbol{\Sigma}_s\Big|_{s=0, z_i=\hat{z}_i}\right)$$

- Features:
  - ▷ it is rather technical (you can get lost in the derivatives of $\{\log \mathsf{det}(\boldsymbol{\Sigma}_s)^{-1/2}\}$ *wrt* s)
  - ▷ however, calculating these terms does not pose a great computational challenge

3.5 Fully Exponential Laplace Approximation

- ▷ an issue with this approximation is that it cannot be used for terms for which $\partial A(\hat{z}_m)/\partial z_m = 0 \Rightarrow$ it cannot be used to calculate the log-likelihood (e.g., to perform LRTs)

3.6 Asymptotic Behaviour of Laplace Estimators

- Let $S(\boldsymbol{\theta})$ true score vector; $\widetilde{S}(\boldsymbol{\theta})$ Laplace-based score vector

$$n^{-1}S(\boldsymbol{\theta}) = n^{-1}\widetilde{S}(\boldsymbol{\theta}) + O\left\{\min(p_i)^{-2}\right\}$$

- Let $\boldsymbol{\theta}_0$ true parameter vector; $\widehat{\boldsymbol{\theta}}$ Laplace-based MLE

$$\begin{aligned} n^{-1}S(\widehat{\boldsymbol{\theta}}) &= n^{-1}S(\boldsymbol{\theta}_0) + n^{-1}H(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + O_p(1) \Rightarrow \\ (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= O_p\left[\max\left\{n^{-1/2}, \min(p_i)^{-2}\right\}\right] \end{aligned}$$

- $\widehat{\boldsymbol{\theta}}$ consistent as both $n, p_i \to \infty$

4 Conclusion

- Results from the similar framework of joint modelling of longitudinal and time-to-event data
  - ▷ Gauss-Hermite requires creating a design matrix of dimensions $N \times h^q$ (N: total sample size; h: quadrature points; q: dimension of integration)
  - ▷ for a data set $h = 3$, $q = 8$ we need 58531×6561 design matrix
  - ▷ One EM iteration
    - * Gauss-Hermite: > 15min
    - * Fully Exponential Laplace Approximation: 12sec

4 Conclusion (cont'd)

- What has been done
  - ▷ theory almost finalized
  - ▷ preliminary R programs written

- What needs to be done
  - ▷ finalize programs
  - ▷ simulation studies

Thank you for your attention!