Generalized Linear Latent Variable Models with Flexible Distribution of Latent Variables

Irina Irincheeva

University of Geneva, Switzerland

Presentation at Swiss Statistics Meeting, October 29, 2009
Joint work with Eva Cantoni and Marc G. Genton

Outline

Latent variables are important in modern science, even everyday life, but cannot be measured directly. Usually we assess them via a linear combination of observable variables.

What?	How to measure?
quality of life	physical and mental health, wealth, employment, education ...
physical health	cholesterol & hemoglobin rates, BMI, chronic disease, eyesight, hearing ...
wealth	expenditures for food, clothes, leisure, owner of car, dishwasher, real estate ...
solvency of a customer	age, permanent employment, revenue, prosecution for debts, color of the car ...

How to construct an optimal linear combination explaining the most of the data?

Generalized Linear Latent Variable Model (GLLVM):

- $h(\mathbf{f})$ is the multivariate density of q latent variables, $\mathbf{f} = (f_1, \ldots, f_q)^T$;
- $g(x_j|\boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{f}),\ j = 1, \ldots, p$ is the conditional density of j-th manifest variable given the latent ones $q < p$, a $p \times 1$ vector $\boldsymbol{\mu}$ and a $p \times q$ matrix $\boldsymbol{\Gamma}$ are parameters of interest;
- Assume that latent variables $\mathbf{f}$ explain all the systematic variability of data i.e. $x_j|\boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{f},\ j = 1, \ldots, p$ are mutually independent. Then the density of random vector $\mathbf{x}$ is

$$g(\mathbf{x}) = \int_{\mathbb{R}^q} \left(\prod_{j=1}^{p} g(x_{ij}|\boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{f}) \right) h(\mathbf{f})\mathrm{d}\mathbf{f}$$

and we can use Maximum likelihood method to estimate parameters.

Example 1: Manifest given latent and latent are normally distributed

$$h(\mathbf{f}) = \frac{1}{(2\pi)^{\frac{q}{2}}} \exp\left\{-\tfrac{1}{2}\mathbf{f}^T\mathbf{f}\right\}, \text{ i.e. } N_q(\mathbf{0}, \mathbf{I})$$

$$g(\mathbf{x}|\mathbf{f}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Psi}|}} \exp\left\{-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu} - \boldsymbol{\Gamma}\mathbf{f})^T\boldsymbol{\Psi}^{-1}(\mathbf{x} - \boldsymbol{\mu} - \boldsymbol{\Gamma}\mathbf{f})\right\},$$

$$\text{i.e. } N_p(\boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{f}, \boldsymbol{\Psi})$$

Result: factor analysis model i.e.

$$\begin{aligned} \mathbf{x} &= \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{f} + \mathbf{u}, \text{ where } \mathbf{f} \sim N_q(\mathbf{0}, \mathbf{I}), \ \mathbf{u} \sim N_p(\mathbf{0}, \boldsymbol{\Psi}), \\ &\quad \mathbf{f} \text{ and } \mathbf{u} \text{ are independent.} \end{aligned}$$

Example 2:

$$x \in \mathbb{R}, \quad f \sim N(0,1), \quad x \mid \mathbf{f} \sim Bernoulli(\theta)$$

so $x \mid \mathbf{f}$ takes values $0, 1$. A possible link function between the expectation of x (which is θ) and $\mu + \gamma f$ is logit:

$$\theta(f) = \frac{\exp\{\mu + \gamma f\}}{1 + \exp\{\mu + \gamma f\}}$$

and the probability mass function of x given f is

$$g(x|f) = \theta(f)^x (1 - \theta(f))^{1-x}.$$

Finally the marginal density of x is

$$g(x) \quad = \quad \int_{\mathbb{R}} g(x|f)h(f)\mathrm{d}f \quad = \quad \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{\exp\{x\mu + x\gamma f - f^2/2\}}{1 + \exp\{\mu + \gamma f\}} \mathrm{d}f$$

The integral does not exist in closed form.

Latent variables were traditionally supposed to be normally distributed due mostly to the Central Limit Theorem. It is not always appropriate.
In our work we propose an approach based on a family of "Semi-NonParametric" (SNP) distributions :

$$h(\mathbf{f}) = P_K^2(\mathbf{f} - \boldsymbol{\tau})\phi(\mathbf{f} \mid \boldsymbol{\tau}, \sigma^2 \mathbf{I}_q), \quad \text{with } \mathbf{f}, \boldsymbol{\tau} \in \mathbb{R}^q,$$

$$P_K(\mathbf{f}) = \sum_{0 \le i_1 + \ldots + i_q \le K} a_{i_1 \ldots i_q} f_1^{i_1} \ldots f_q^{i_q},$$

$a_{i_1 \ldots i_q}$ are such that $h(\mathbf{f})$ is a valid density function,

$\phi_q(\mathbf{f} \mid \boldsymbol{\tau}, \sigma^2 \mathbf{I}_q)$ is the density of $N_q(\boldsymbol{\tau}, \sigma^2 \mathbf{I}_q)$.

This densities can approximate any smooth enough density arbitrarily close (Gallant and Nychka (1987)). This approach was already used by Chen, Zhang, and Davidian (2002) for random effects in Generalized Linear Mixed Models.

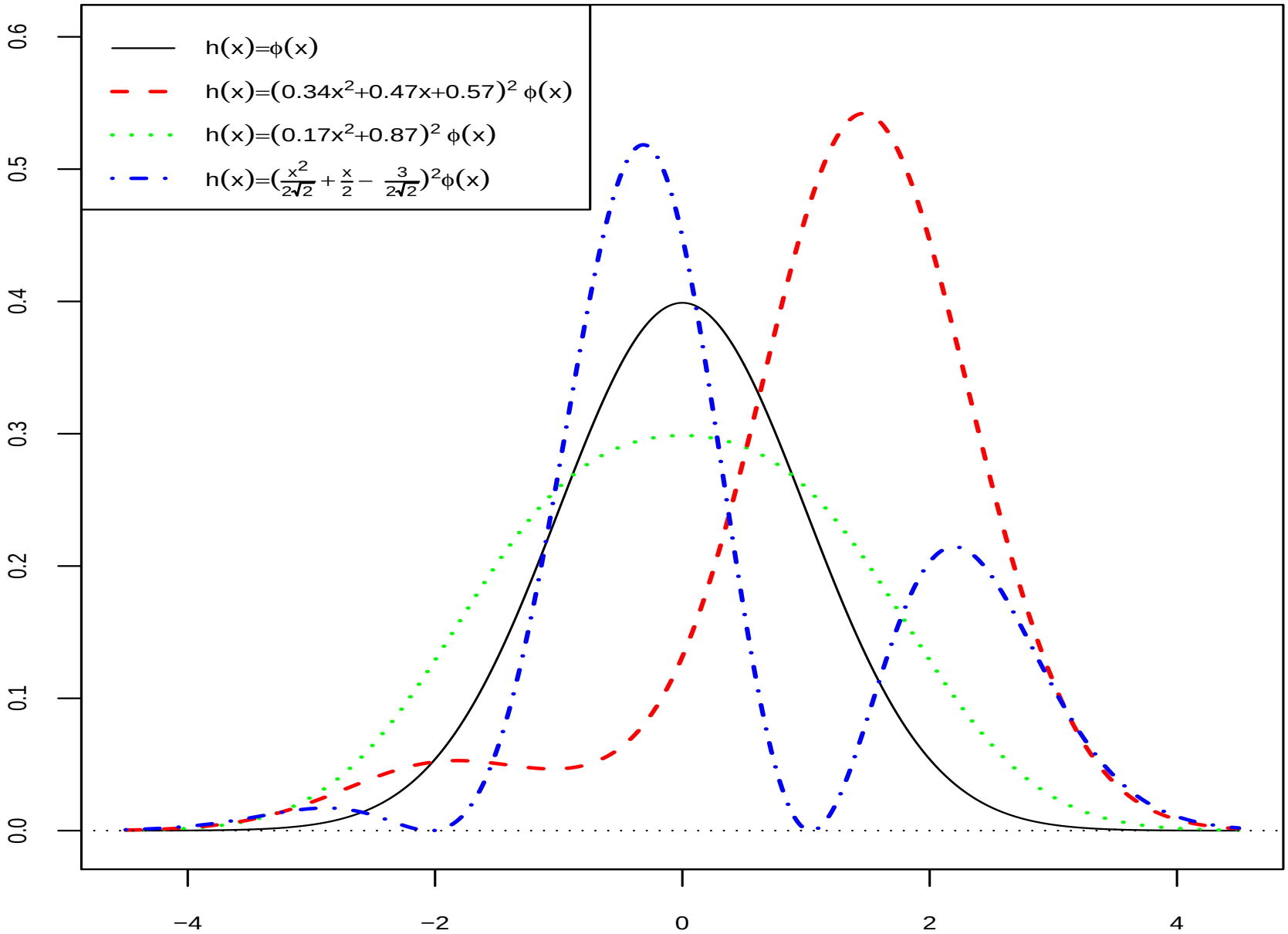


Figure: Some particular cases for $K = 1$ (SNP1) and 2 (SNP2) of univariate densities of the form $h(f) = P_K^2(f)\phi(f)$.

If

$$h(\mathbf{f}) = P_K^2(\mathbf{f})\phi(\mathbf{f} \mid \mathbf{0}, \mathbf{I}),$$

and

$$g(\mathbf{x}|\mathbf{f}) = \phi(\mathbf{x} \mid \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{f}, \boldsymbol{\Psi})$$

then $g(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Psi})$ has closed form for any K!
We implemented the estimation by

- SNP1 ML, i.e. assuming that latent density is $(a_0 + a_1 f)^2 \phi(f)$
- SNP2 ML, i.e. assuming that latent density is $(a_0 + a_1 f + a_2 f^2)^2 \phi(f)$.

Analytical gradient and hessian are used to boost the optimization which is very sensitive to the initial values.

We explore the properties of proposed estimators on 200 simulated samples of size 200 from the model: $\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} f + \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$ where

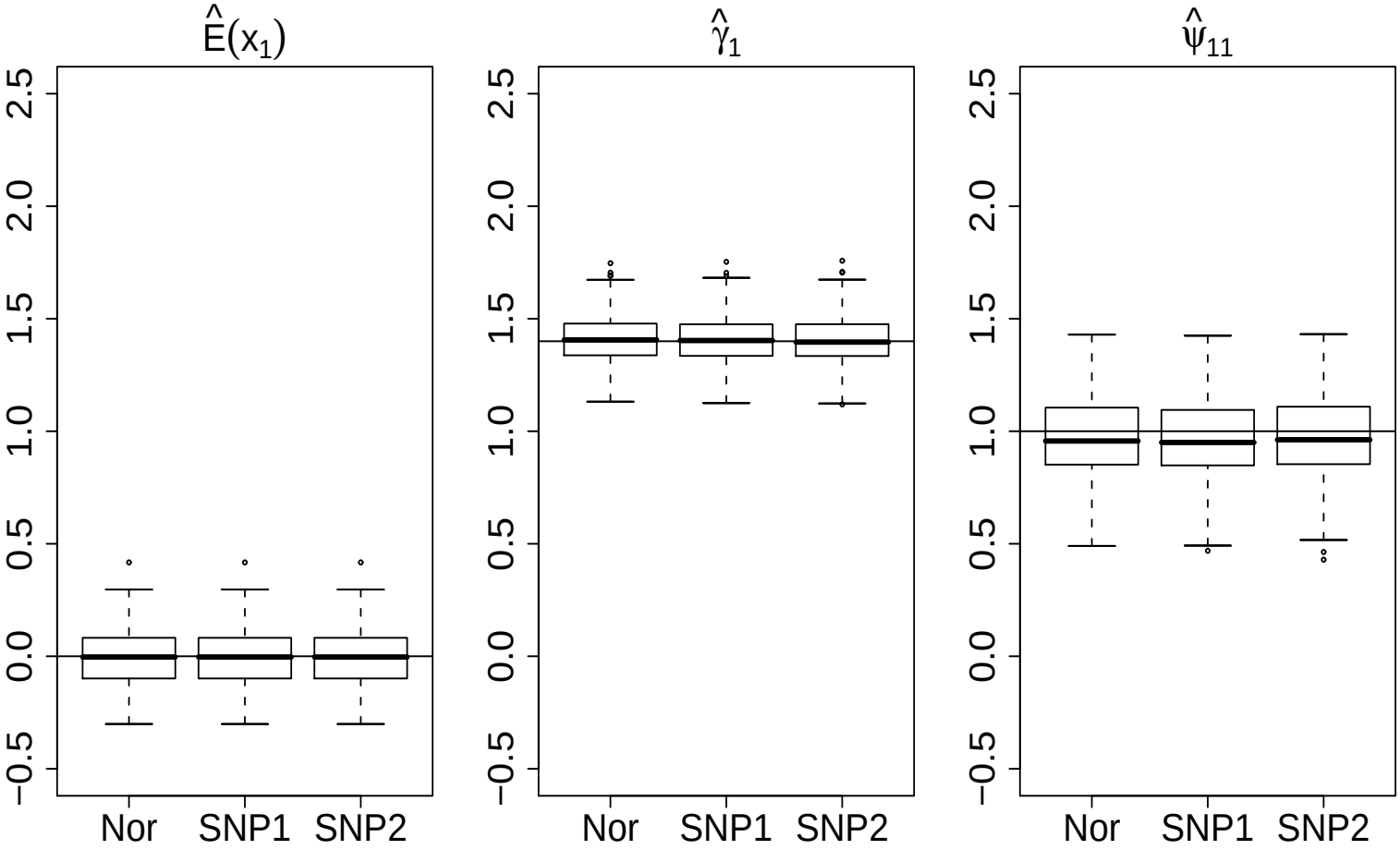
$$u_i \sim N(0, \psi_i), \quad \psi_1 = \psi_2 = \psi_3 = 1,$$

$$\gamma_1 = 1.4, \ \gamma_2 = 1.8, \ \gamma_3 = -1,$$

and we generate f from 3 different distributions:

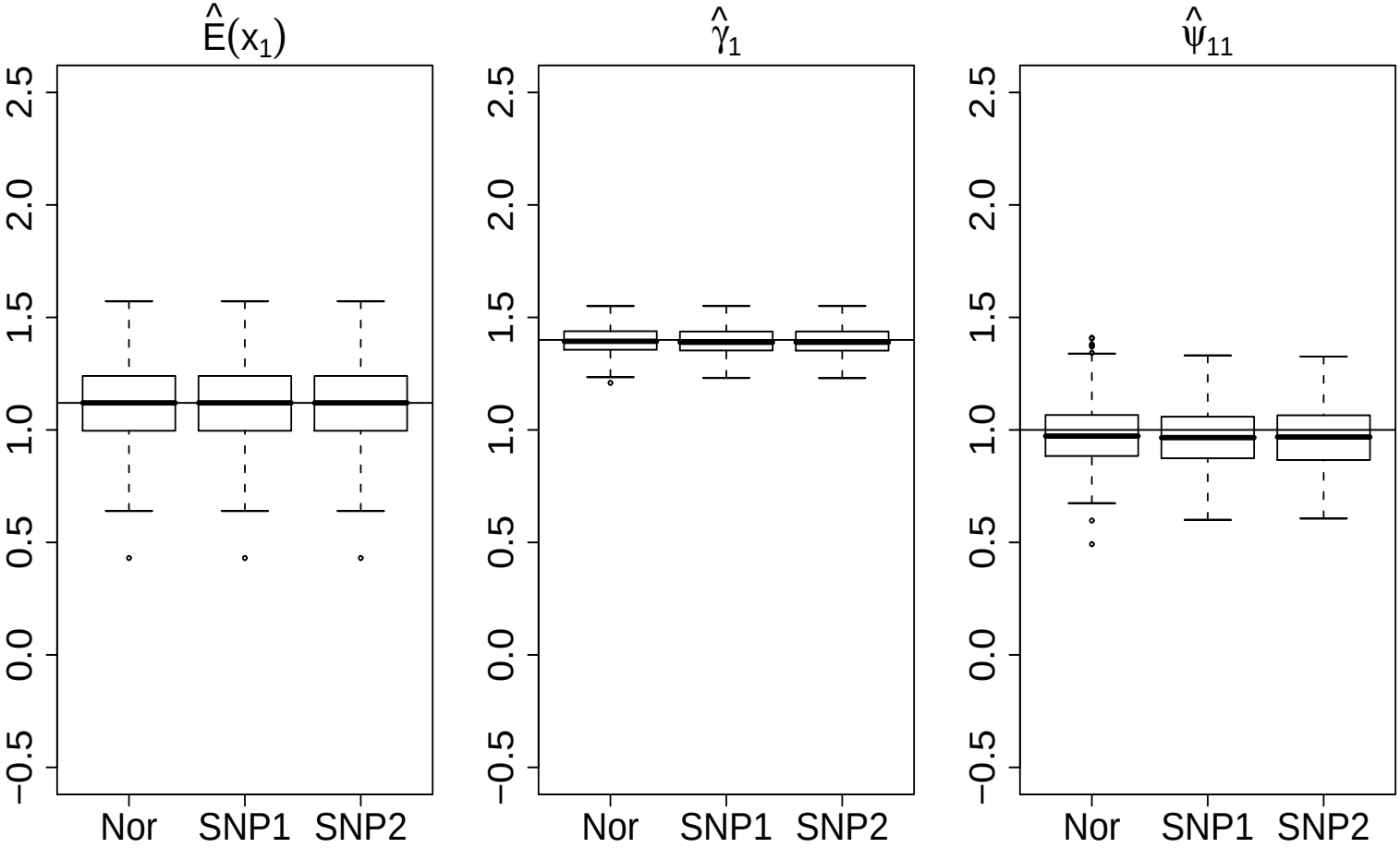
- standard normal distribution $N(0, 1)$;
- mixture of normals with distribution $0.7N(2, 1) + 0.3N(-2, 0.25)$
- SNP2 distribution with density
  $\left(-\frac{1}{\sqrt{2}} \cos 0.7 + f \sin 0.7 + f^2 \frac{1}{\sqrt{2}} \cos 0.7\right)^2 \phi(f)$

Standard normal latent variable



Mixture of normals latent variable

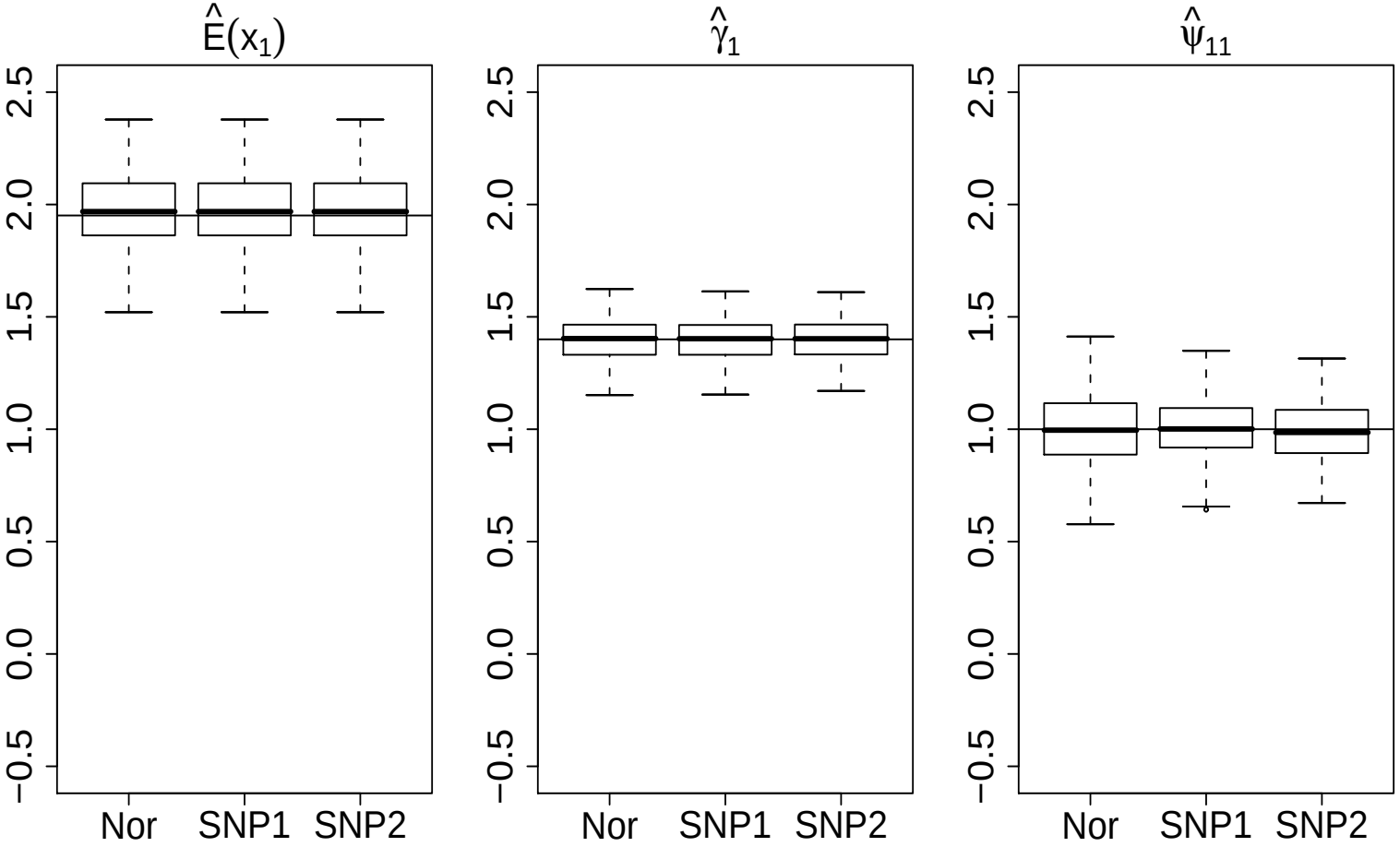
Latent variable is generated from the mixture of normals
$0.7N(2,1) + 0.3N(-2, 0.25)$



SNP2 latent variable

Latent variable is generated from the SNP2 density with distribution

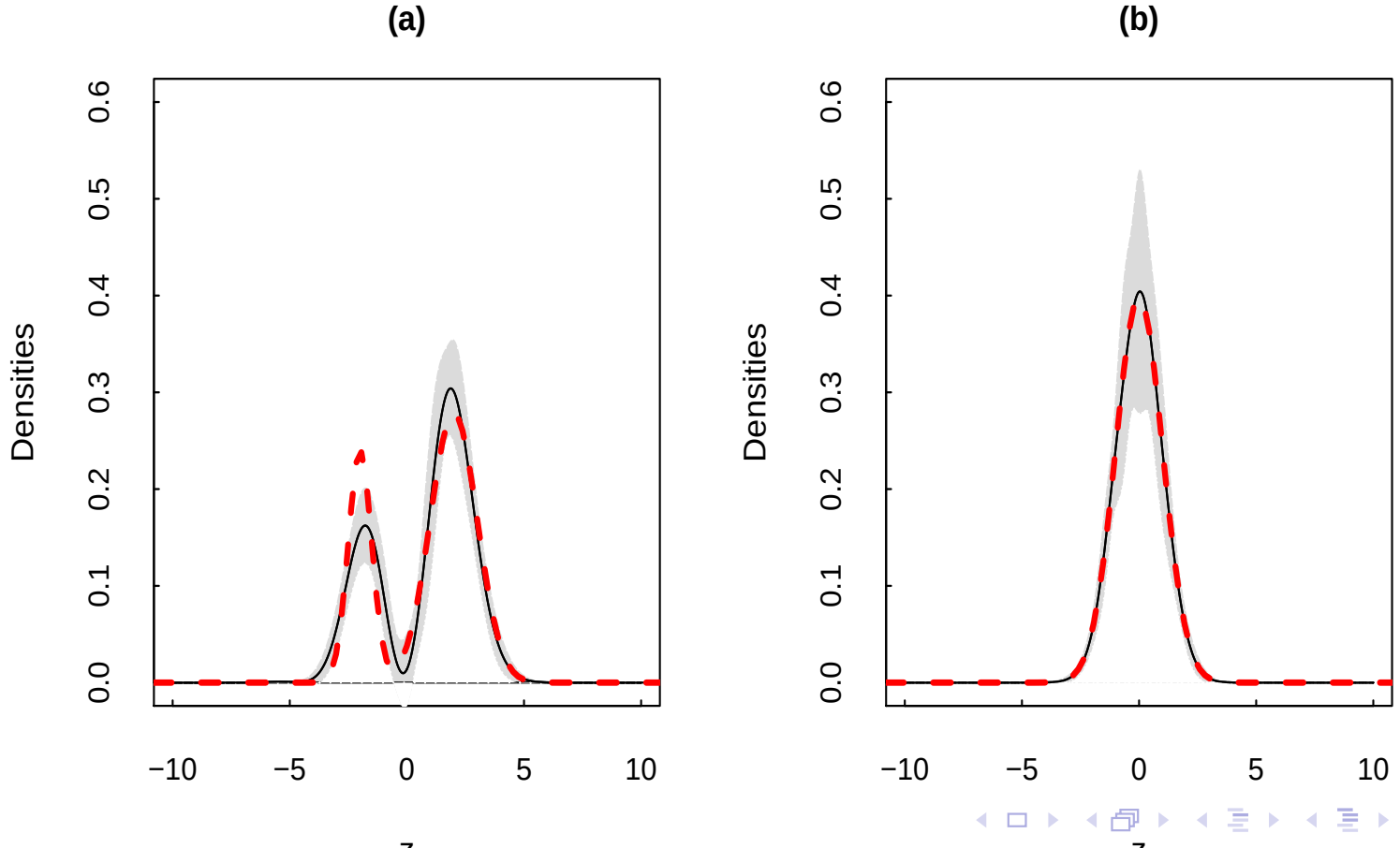
$$\left(-\frac{1}{\sqrt{2}}\cos 0.7 + f \sin 0.7 + f^2 \frac{1}{\sqrt{2}}\cos 0.7\right)^2 \phi(f)$$



Surprisingly, the FA MLE of loadings and uniquenesses are not sensitive at all to the wrong specification of the latent variable distribution. But our approach allows to visualize the latent variable distribution.

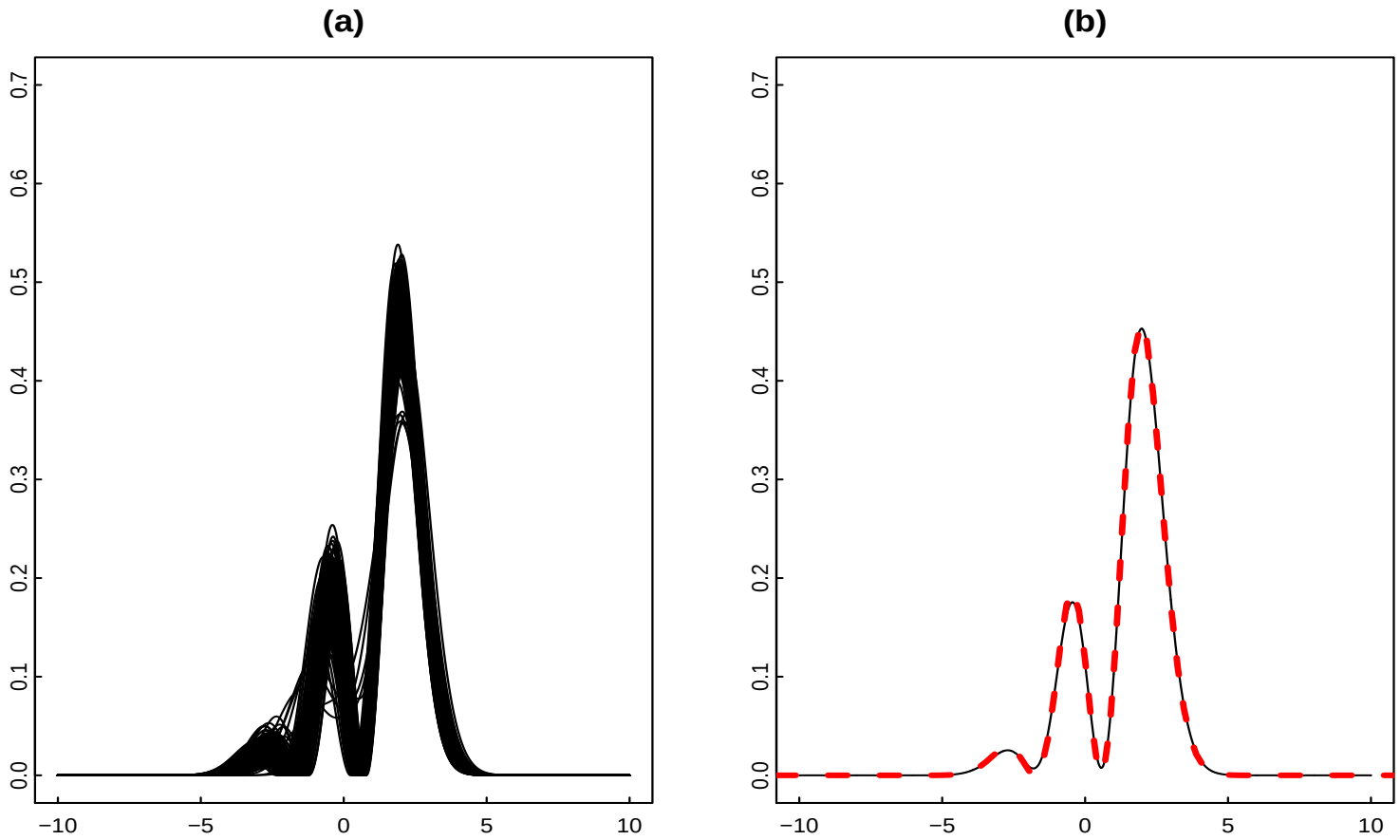
Estimated densities 1

Estimated latent densities when the true is normal and mixture of normals. **(a), (b)** Solid line is the average of estimated densities for fits preferred by HQ, shaded area is the point wise estimated confidence envelope, dashed red line is the true density.



Estimated densities 2

Simulations results based on 200 data sets with SNP2 latent: **(a)** Densities estimated by SNP2. **(b)** Solid curve is the average of SNP2 estimated densities, dashed red curve is the true density.



Discussion

- Remarkable "robustness" of FA to the wrong specification of the latent variable distribution. Similar studies in mixed linear models confirm this result.
- This "robustness" does not hold in cases when distribution of manifest given the latent is discontinuous (Bernoulli for example). We are investigating this case.
- Our approach offer a deeper insight on the behavior of latent variables. Indeed the non-normality of the estimated latent density can indicate presence of outliers, non-linearity, heterogeneity of population or just inconvenience of normally distributed latent.
- Different fits (Nor, SNP1, SNP2) can be compared using information criterions such as AIC, BIC or HQ.

Chen, J., Zhang, D., & Davidian, M. (2002). A Monte-Carlo EM Algorithm for generalized mixed models with flexible random effects distribution. *Biostatistics*, *3*(3), 347-360.

Gallant, A. R., & Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, *55*(2), 363-390.