Slide 1

Alejandra Avalos Pacheco, "Bayesian Factor Regression Analysis in Heterogeneous High-dim Biological Data"

Peter Müller

Summary: super nice paper; multi-study factor analysis

$$x_i = \phi f_i + \theta v_i + \beta b_i + e_i$$

or

$$x_{si} = \phi f_{si} + \theta v_{si} + \beta_s + e_{si}$$

for patients in studies $s = 1, 2$.

Factor loadings: ϕ, common across all batches (studies), sparse

Regression: θv_i, regression on covariates

Batch effects: b_i, design vector for (additive) batch effects

Multiplicative batch effects: $e_i \mid i \in$ batch s $\sim N(0, \text{diag}(\tau_{1s}^{-1}, \ldots, \tau_{ps}^{-1}))$, with batch-specific variance τ_{js}^{-1}.

Slide 2

Highlights

Multi-study factor analysis: builds on approaches like deVito, Bellio, Trippa & Parmigiani (2018 Bmcs), but importantly adds *sparsity* in ϕ, NLP's and joint inference for batch effects (rather than pre-processing)

Spike & slab on ϕ_{jk}: NLP on non-zero ϕ_{jk} avoids prob mass at 0, making it easier to interpret non-zero values ($\gamma_{jk} = 1$)

IBP: interesting prior on γ_{jk} (which could easily extend to random # factors).

Computation: efficient EM style algorithm

Slide 3

Questions

Batch effects: Define batch effects as non-biological variation when *"data are generated under different experimental conditions."* But you still assume the same common factor model part?

Question: What if the measurements are different? Study design or patient populations are different? Or covariate vectors (v_i) are different dimension or variable definition?

For example, two studies of the same treatment, but definition of the endpoint, different eligibility criteria, randomized clinical trial vs. observational data etc.

Slide 4

Questions

Sampling model: factor analysis can recover common underlying structure. This works with normal sampling model and linear structure.

Question: But many effects are surely not linear? For example, single cell data requires more complex models (as in BASiCS in Vallejos et al, 2015).

Slide 5

Questions

Factor analysis: Is "factor analysis" with the normal linear structure the right way to formalize the search for underlying common biologic processes? How can one interpret ϕf_i?

Question: Can we

- interpret columns of ϕ as identifying underlying processes (cell or patient sub-populations), and
- f_{ik} as relative proportions *and* effects of the k-th underlying xx?

The sparse prior on ϕ helps a bit, but the additive decomposition into effects of different underlying factors is not obvious.
For example, in omics data, raw data are heavily preprocessed. Decomposition on raw data (counts etc.) $\not\Rightarrow$ decomposition on transformed (normalized etc.) data.

Slide 6

Questions

Alternative model: How about models that are more closely aligned with the interpretation of "col of $\phi \leftrightarrow$ biologic factors:

Φ = binary matrix to identify subsets

and then use whatever sampling model is needed. The normal linear model is a bit very special!

Funnily, this implicitly shows up in the IBP prior for non-"zero" Φ_{jk}. The IBP can be interpreted as a prior on a family of subsets (the k-th subset defined by the 0/1 indicators in the k-th column of Φ).

Question: Why do you chose the IBP prior for γ_{jk}? Did you already intend it as a prior on random subsets?

Slide 7

Prior vs. preferences: Using the NLP is a clever prior choice to favor interpretable results - avoids clinically meaningless positives, and helps the interpretation of Φ.

Question: Using the NLP, are we are confounding

- estimation (by introducing prior information on the ϕ_{jk}) on one hand,
- a decision problem (by stating preferences for non-zero ϕ_{jk})?

Would it be better to separate the two aspects?

Slide 8

Adjusting for study differences: The model can fit data across different studies (as in your examples), accomodating study-specific effects.

Question: can I use your model to adjust for study-specific effects? I.e., as part of the inference could i adjust the data such that the two studies look the same?

Question: related question, finding only insignificant study-specific effects, could I use that as an operational proof of equivalent study populations?

Can you guess in which application this becomes useful :-)

Slide 9

Thanks – great paper to use several clever tricks & techniques to address an important problem.
Great step forward from existing literature. And actually feasible!