Inference in generalized bilinear models

Jeff Miller

Joint work with Scott L. Carter

Harvard T.H. Chan School of Public Health
Department of Biostatistics

New England Statistics Symposium (NESS)
University of Connecticut, May 24, 2022

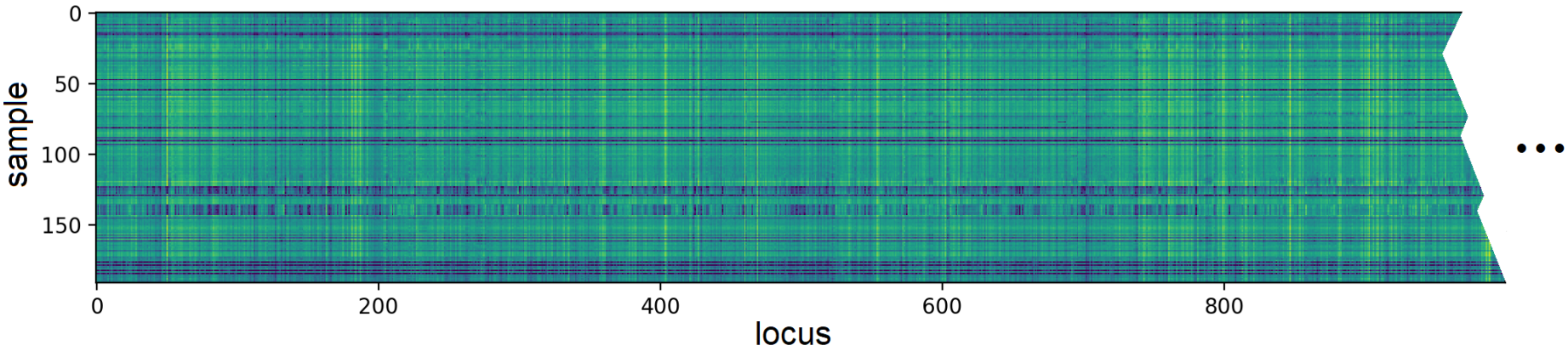
Preprint: https://arxiv.org/abs/2010.04896
Slides: http://jwmi.github.io/talks/ness2022.pdf

Background

Modern high-throughput sequencing yields large matrices of counts.

- Copy ratio estimation in cancer genomics
  - whole-exome or whole-genome sequencing data
- Copy number variation in genetics
  - whole-exome or whole-genome sequencing data
- Gene expression analysis in biology/medicine
  - RNA-seq data for transcript abundance

log counts for a whole-exome seq data set of 191 samples $\times$ 171523 loci



Background

- Latent factor models are widely used to discover and adjust for hidden variation in these applications and many others.
- Estimation and inference in latent factor models is challenging.
- Consequently, most methods do not fully account for uncertainty in the latent factors, which can lead to miscalibrated inferences such as overconfident p-values.

This talk

- Generalized bilinear models (GBMs) are a flexible extension of generalized linear models (GLMs) to include latent factors as well as row covariates, column covariates, and interactions.
- We propose fast and accurate methods for GBM estimation and inference (i.e., uncertainty quantification).
- We introduce *delta propagation*, a novel technique for propagating uncertainty among model components using the delta method.
- We present simulation studies assessing performance.
- We apply GBMs to copy ratio estimation and RNA-seq analysis.

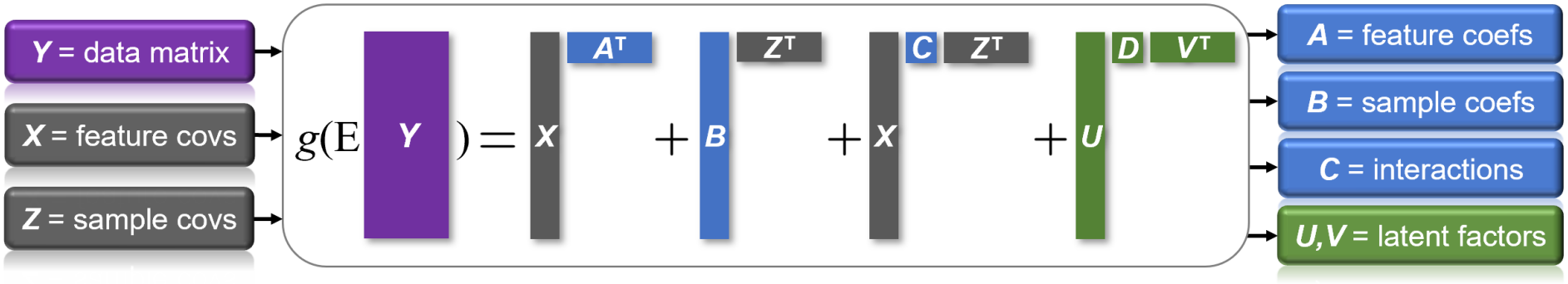
Outline

1. Generalized bilinear models (GBMs)
2. Previous work
3. Estimation
4. Inference (uncertainty quantification)
5. Application
   - RNA-seq gene expression analysis

Outline

1. Generalized bilinear models (GBMs)
2. Previous work
3. Estimation
4. Inference (uncertainty quantification)
5. Application
   - RNA-seq gene expression analysis

Generalized bilinear models (GBMs)



- Suppose the data matrix $\boldsymbol{Y} = (Y_{ij}) \in \mathbb{R}^{I \times J}$ satisfies

$$g(\mathrm{E}(\boldsymbol{Y})) = XA^{\mathrm{T}} + BZ^{\mathrm{T}} + XCZ^{\mathrm{T}} + UDV^{\mathrm{T}}$$

  where the link function g is applied element-wise.

- We refer to this as a *generalized bilinear model* (Choulakian, 1996).

- The "bilinear" part UDV^{T} is a low-rank matrix that captures latent effects due, for example, to unobserved covariates such as batch.

Outcome distributions

- We consider discrete exponential dispersion families (EDFs).
- Specifically, we suppose $Y_{ij} \sim f(y \mid \theta_{ij}, r_{ij})$ where

$$f(y \mid \theta, r) = \exp(\theta y - r\kappa(\theta))h(y, r).$$

- For any discrete EDF,

$$\mu = \mathrm{E}(Y) = r\kappa'(\theta)$$
$$\sigma^2 = \mathrm{Var}(Y) = r\kappa''(\theta).$$

- For sequencing data, we focus on negative binomial outcomes, which is a special case of discrete EDF.
- We parametrize the dispersions as $1/r_{ij} = \exp(s_i + t_j + \omega)$.

Outline

1. Generalized bilinear models (GBMs)
2. Previous work
3. Estimation
4. Inference (uncertainty quantification)
5. Application
   - RNA-seq gene expression analysis

Previous work: Normal bilinear models with covariates

- Consider the following special case:

$$\boldsymbol{Y} = XA^{\mathsf{T}} + BZ^{\mathsf{T}} + XCZ^{\mathsf{T}} + UDV^{\mathsf{T}} + \boldsymbol{\varepsilon}$$

  where $\boldsymbol{\varepsilon}$ is a matrix of residuals with $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_{ij})$.

- Work on this model was inspired by Tukey (1962), who suggested combining regression with factor analysis.

- Estimation for this model, assuming $\sigma^2_{ij} = \sigma^2$:

  Gabriel (1978), Takane and Shibayama (1991).

- Hypothesis testing and confidence regions, assuming $\sigma^2_{ij} = \sigma^2$:

  Perry and Pillai (2013) show how to perform inference for univariate linear projections of A and B.

Previous work: GBMs with covariates

- Consider the general case:

$$g(\mathrm{E}(\boldsymbol{Y})) = XA^{\mathsf{T}} + BZ^{\mathsf{T}} + XCZ^{\mathsf{T}} + UDV^{\mathsf{T}}.$$

- Previous authors have considered models of this form:

  Choulakian (1996), Gabriel (1998), de Falguerolles (2000), Townes (2019).

- Townes (2019) develops a fast estimation algorithm using diagonal approximations to Fisher scoring updates for ℓ_2-penalized estimation.

- Limitations of previous work:
  - uncertainty quantification is not addressed,
  - a single common dispersion parameter is assumed, and
  - identifiability constraints are not explicitly enforced during estimation.

Outline

1. Generalized bilinear models (GBMs)
2. Previous work
3. Estimation
4. Inference (uncertainty quantification)
5. Application
   - RNA-seq gene expression analysis

Estimation algorithm

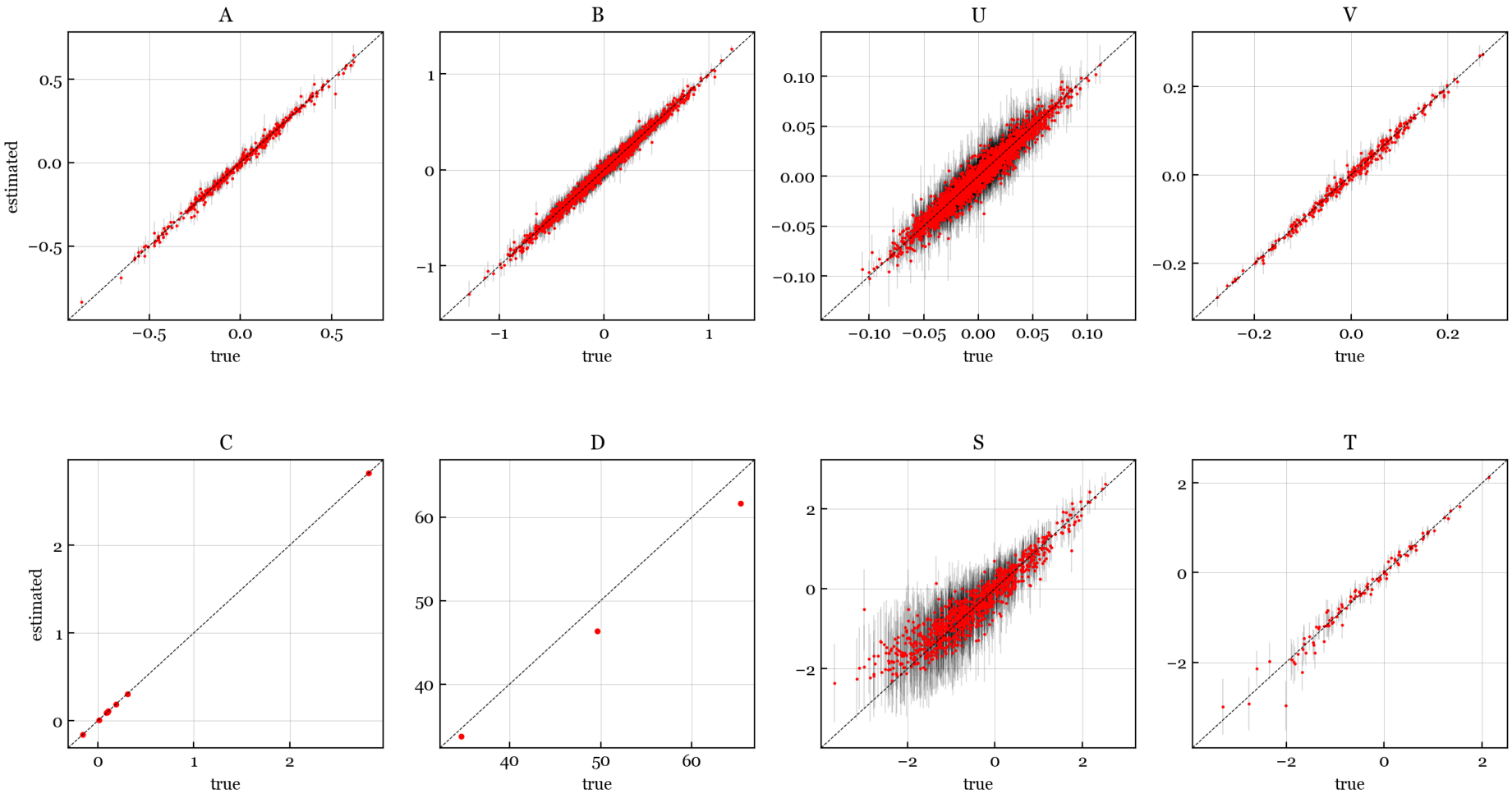
- We provide an algorithm for *maximum a posteriori* GBM estimation that extends previous work by:
  - estimating row- and column-specific dispersion parameters,
  - improving numerical stability, and
  - explicitly enforcing identifiability constraints during estimation.

- Basic idea: Iteratively cycle through the components of the model, updating each in turn using an optimization-projection step.

- "Optimization-projection" = unconstrained optimization step and a likelihood-preserving projection onto the constrained parameter space.

Estimation: Typical example

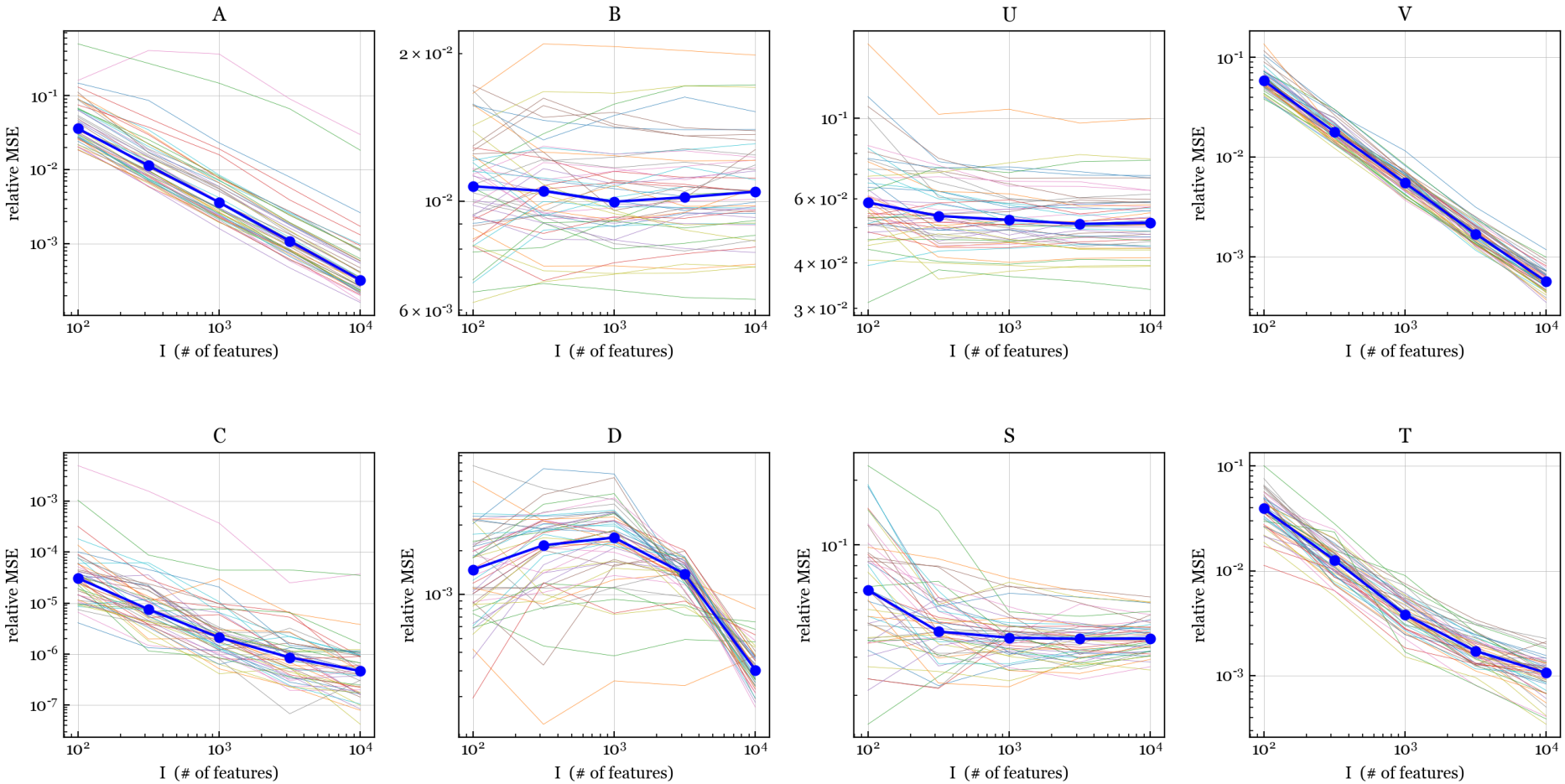
Scatterplots of estimated versus true parameters for a typical simulated data matrix (`NB/Normal/Normal`, 1000 rows, 100 cols, 4 feature covs, 2 samples covs, and 3 factors)



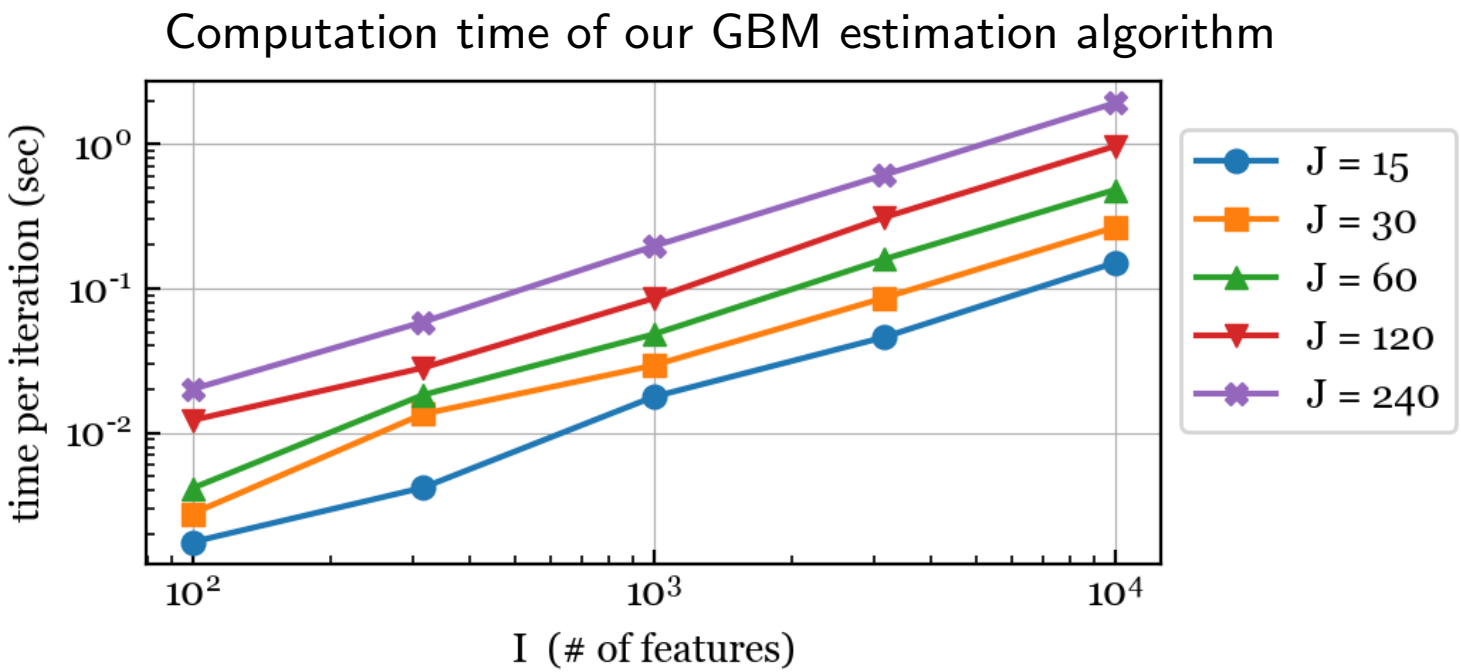
Estimation: Error tends to zero with increasing data

Relative mean-squared error between estimated and true parameter values

(50 runs of `NB/Normal/Normal`, 100 cols, 4 feature covs, 2 samples covs, and 3 factors)



Estimation: Computation time is linear in size of matrix



- Computation time grows linearly with I (# rows) and J (# cols).
- Each dot is the average over 10 runs of the `NB/Normal/Normal` scheme with 4 feature covs, 2 samples covs, and 3 factors.
- The empirical results agree with the theory.

Outline

1. Generalized bilinear models (GBMs)
2. Previous work
3. Estimation
4. Inference (uncertainty quantification)
5. Application
   - RNA-seq gene expression analysis

Inference (uncertainty quantification)

- Most latent factor methods do not fully account for uncertainty.
- To remove batch effects in gene expression, several methods estimate UDV^{T} and then treat V as known, handling uncertainty only in U.

  Leek and Storey (2007, 2008), Sun et al. (2012), Risso et al. (2014)
- CNV detection methods often fit UDV^{T} and just subtract it off.

  Fromer et al. (2012), Krumm et al. (2012), Jiang et al. (2015)
- Bayesian inference provides full uncertainty quantification, but MCMC is slow in large parameter spaces with strong dependencies.
- Variational Bayes is faster, but relies on factorized approximations that tend to underestimate uncertainty.

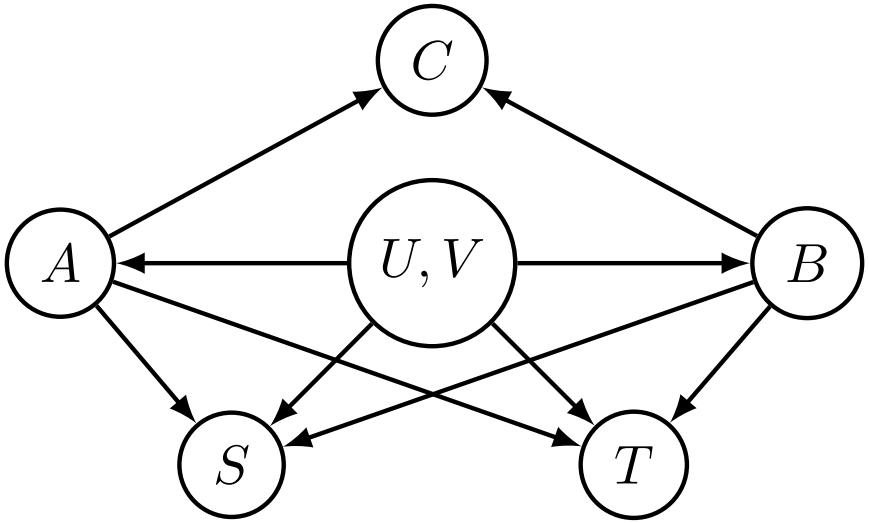
  Stegle et al. (2010), Buettner et al. (2017), Babadi et al. (2018)

Inference: Novel method – "delta propagation"

- We provide a fast, accurate method for GBM uncertainty quantification.
- In particular, we introduce *delta propagation*, a general technique for propagating uncertainty among model components using the delta method.
- Delta propagation can be done analytically using closed-form expressions involving the gradient and the Fisher information.

Inference: Outline of GBM inference algorithm

Diagram of uncertainty propagation scheme for GBM inference

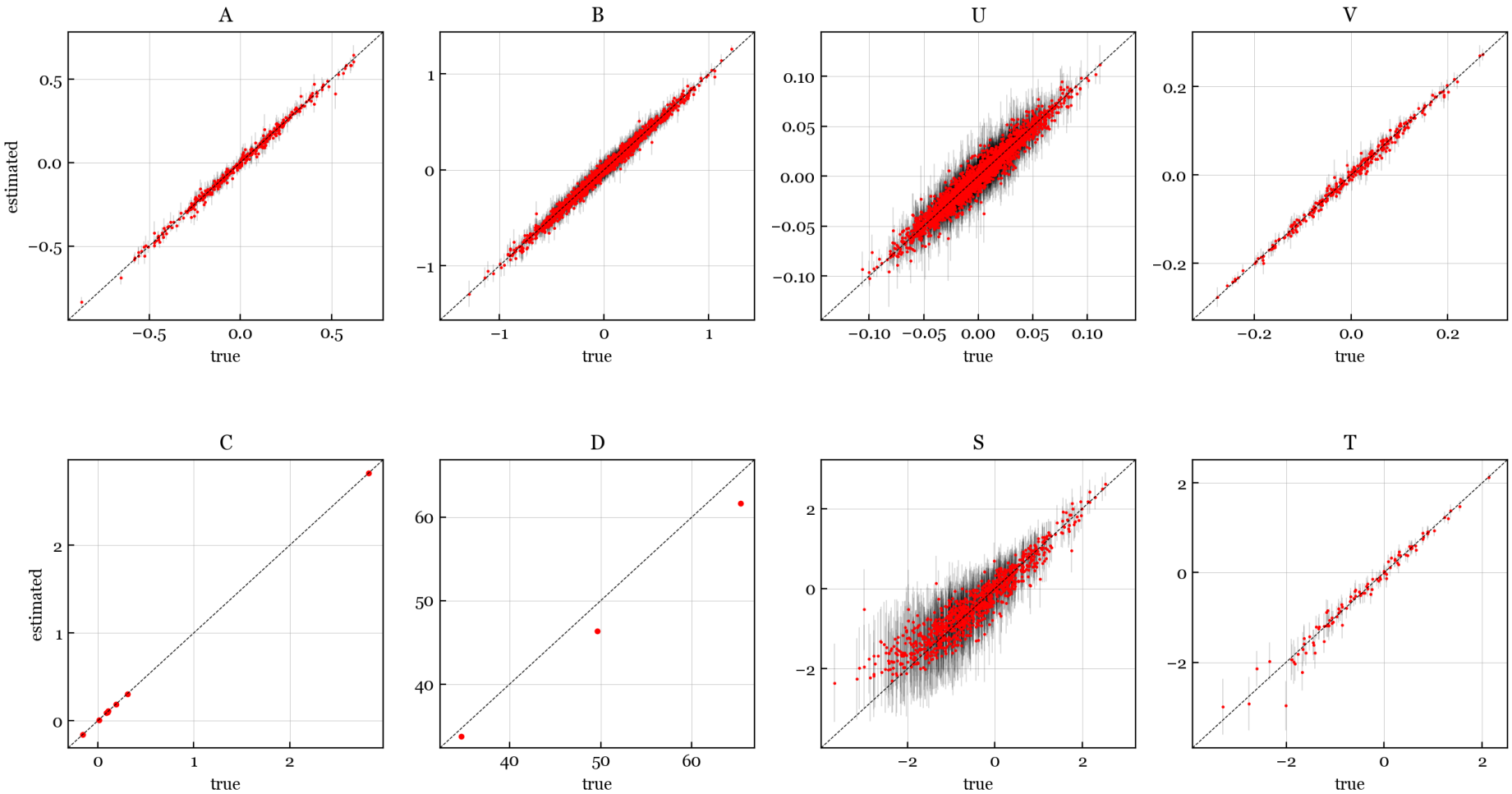


Outline:

1. Compute conditional uncertainty for each parameter matrix/vector.
2. Compute joint uncertainty in (U, V) accounting for constraints.
3. Propagate uncertainty between components using delta propagation.
4. Compute approximate standard errors.

Inference: Typical example

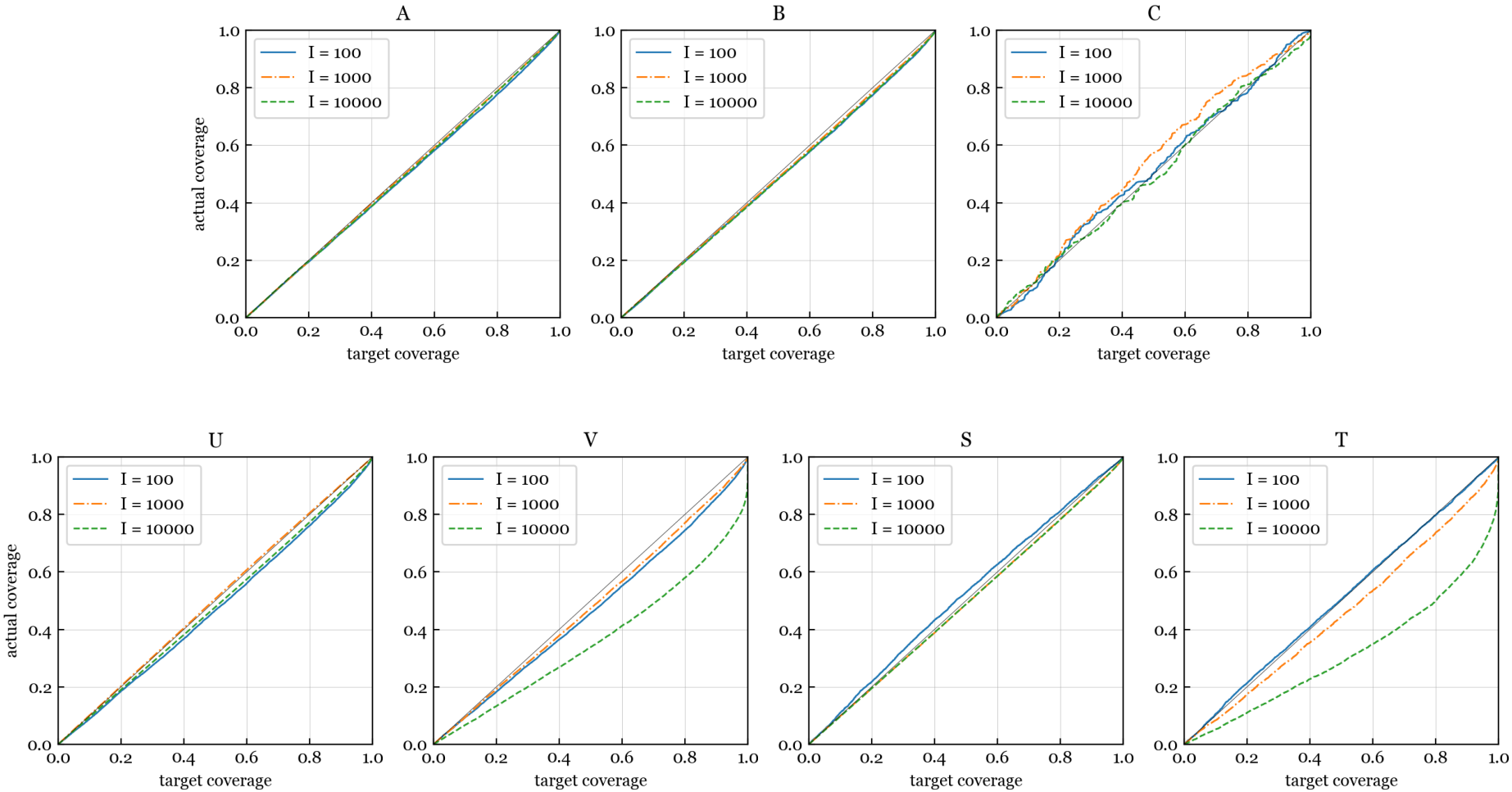
Scatterplots of estimated versus true parameters for a typical simulated data matrix (`NB/Normal/Normal`, 1000 rows, 100 cols, 4 feature covs, 2 samples covs, and 3 factors)



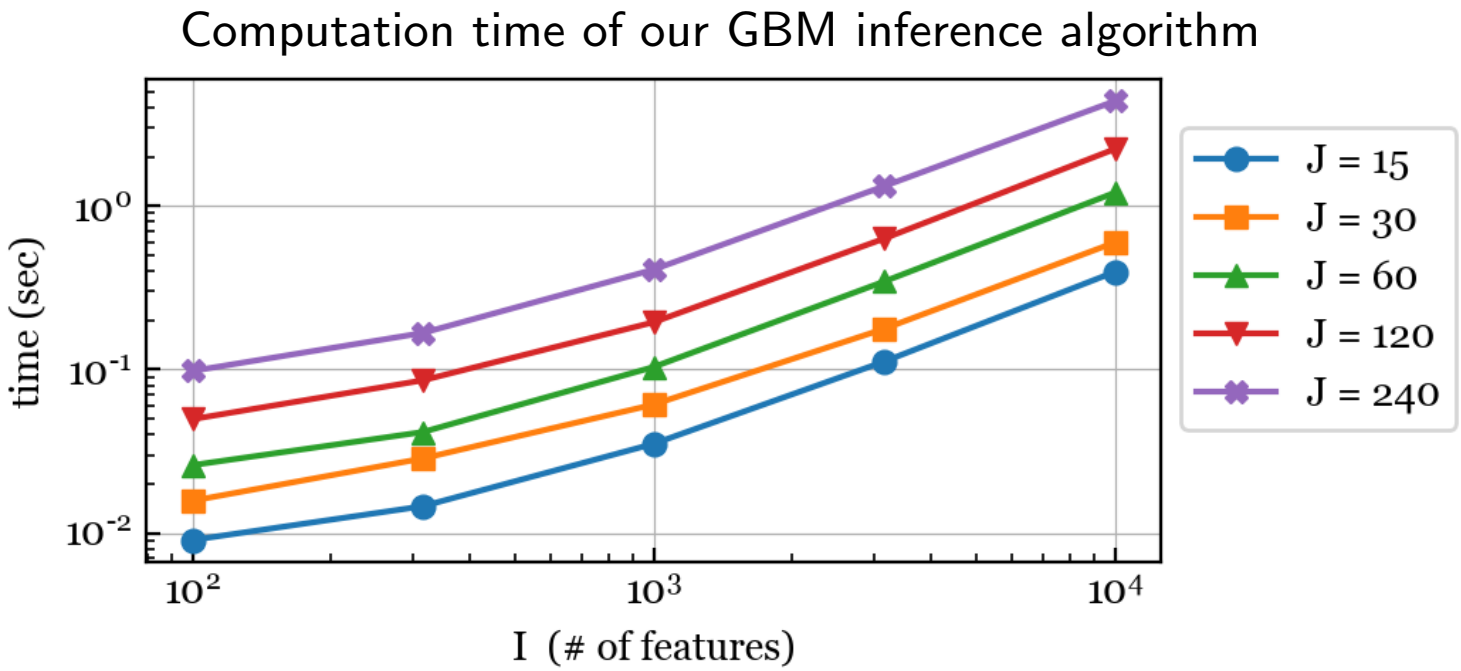
Inference: Coverage is good for most params of interest

Coverage of confidence intervals for the entries of each parameter matrix/vector

(50 runs of `NB/Normal/Normal`, 100 cols, 4 feature covs, 2 samples covs, and 3 factors)



Inference: Empirical assessment of computation time



- Theory indicates that computation time is linear in I (# rows) and quadratic in J (# cols).
- Thus, as I increases, the curves should become linear in I.
- Each dot is the average over 10 runs of the `NB/Normal/Normal` scheme with 4 feature covs, 2 samples covs, and 3 factors.

Outline

1. Generalized bilinear models (GBMs)
2. Previous work
3. Estimation
4. Inference (uncertainty quantification)
5. Application
   - RNA-seq gene expression analysis

RNA-seq: Analyzing GTEx data for aging-related genes

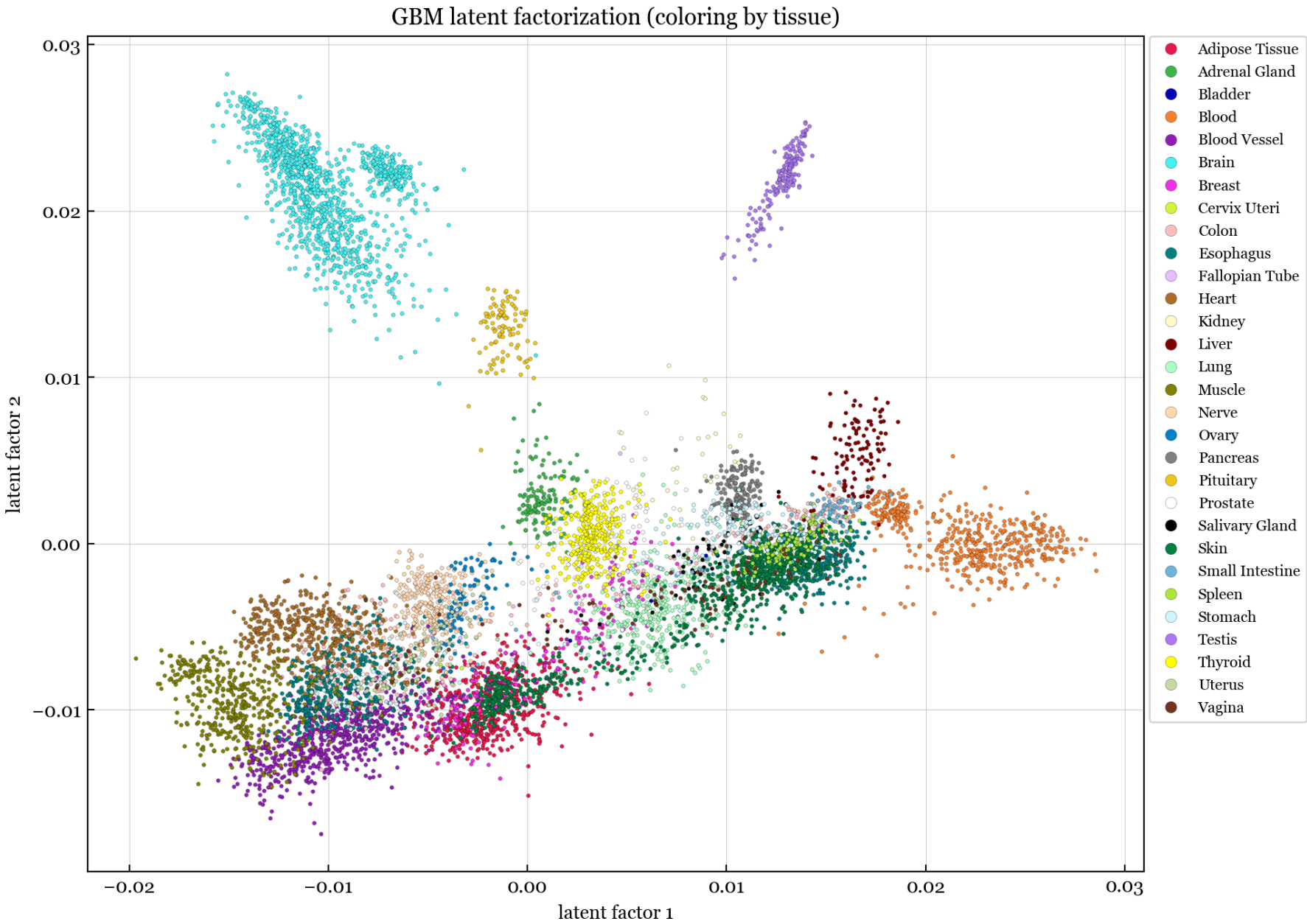
- We consider RNA-seq data from the Genotype-Tissue Expression (GTEx) project (Melé et al., 2015).
- 8,551 samples from 30 tissues in the human body, from 544 subjects.
- We apply the GBM to find genes whose expression changes with age, adjusting for technical biases.

RNA-seq: PCA of GTEx data using log-transformed TPMs

RNA-seq: Visualizing GTEx data using a GBM

- Similar to PCA, we can use the GBM to visualize high-dimensional data by plotting the V matrix.
- First, we take a random subset of 5,000 genes and fit a negative binomial GBM with:
  - two latent factors,
  - no sample covariates, and
  - $\log(\texttt{length}_i)$, gc_i, and $(\text{gc}_i - \overline{\text{gc}})^2$ as gene covariates.
- Model dimensions: $I = 5{,}000$, $J = 8{,}551$, $K = 4$, $L = 1$, and $M = 2$.

RNA-seq: Visualizing GTEx data using a GBM



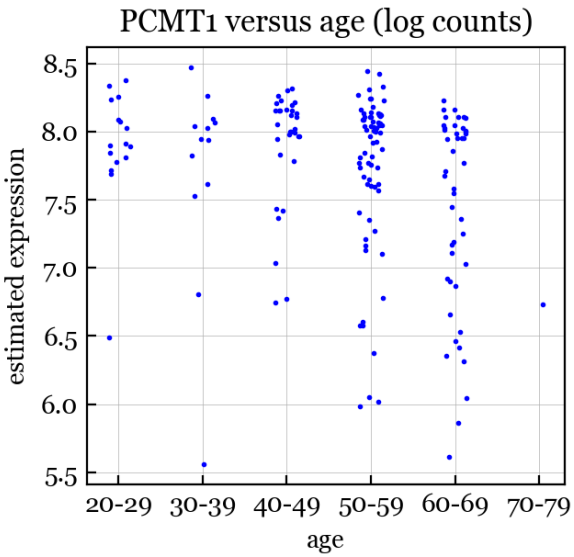
RNA-seq: Analyzing GTEx data for aging-related genes

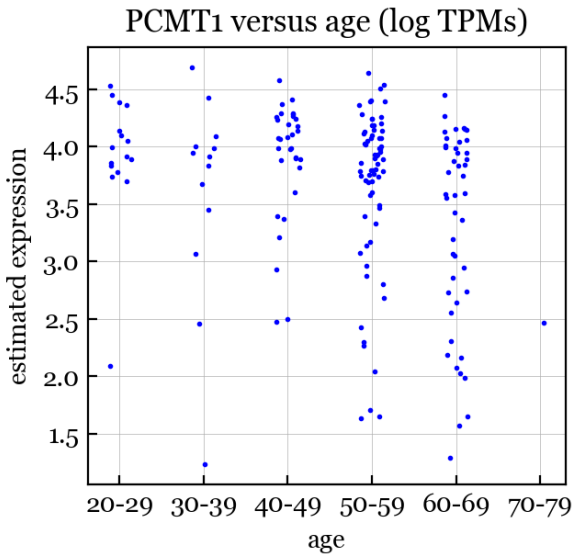
- To find aging-related genes, we add subject age as a sample covariate.
- For illustration, we present results for the "Heart - Left Ventricle" subtissue (Heart-LV).
- We ran the GBM on the 176 Heart-LV samples in the test set, using:
  - the 19,853 genes with nonzero median across these samples,
  - gene covariates: $\log(\texttt{length}_i)$, $\texttt{gc}_i$, and $(\texttt{gc}_i - \overline{\texttt{gc}})^2$,
  - sample covariates: `smexncrt` (exonic rate) and age (subject age),
  - 3 latent factors.
- This choice of subtissue and model was based on the exploratory phase.

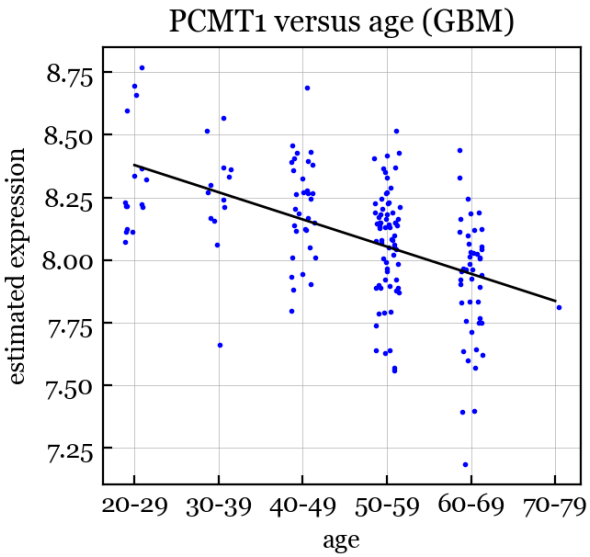
RNA-seq: Analyzing GTEx data for aging-related genes

- In this GBM, each gene has a coefficient describing how its expression changes with age.
- Using our GBM inference algorithm, we compute a p-value for each gene to test whether this coefficient is nonzero.
- 2,444 genes were significantly associated with age in Heart-LV, controlling Type I error at 0.05 using Bonferroni.
- For comparison, simple linear regression on the log-transformed TPMs yields only 1 significant gene.
- This indicates that the GBM has much greater power than a simple standard approach.

RNA-seq: Expression of the top aging-related gene







- The top GBM hit for Heart-LV is PCMT1 (p-value $= 1.1 \times 10^{-47}$).
- PCMT1 is involved in the repair and degradation of damaged proteins, and is a well-known aging gene (Tacutu et al., 2018).
- GBM-estimated expression of PCMT1 exhibits a clear downward linear trend with age.
- The log TPMs for PCMT1 are noisier and the trend is much less clear.

RNA-seq: Top age-related GO terms (Biological Process)

- To test for enrichment of Gene Ontology (GO) term gene sets, we run DAVID on the top 1000 GBM hits for Heart-LV.
- These results are highly consistent with known aging biology (López-Otín et al., 2013).

GO term ID	Description	Count	p-value	Benjamini
GO:0098609	cell-cell adhesion	48	5.1e-12	1.5e-08
GO:0006418	tRNA aminoacylation for protein translation	16	1.4e-09	2.0e-06
GO:0006099	tricarboxylic acid cycle	12	3.7e-07	3.6e-04
GO:1904871	positive regulation of protein localization to Cajal body	7	1.1e-06	6.1e-04
GO:1904851	positive regulation of establishment of protein localization to telomere	7	1.1e-06	6.1e-04
GO:0006607	NLS-bearing protein import into nucleus	10	1.3e-06	6.2e-04
GO:0006914	autophagy	22	1.8e-05	7.6e-03
GO:0016192	vesicle-mediated transport	24	2.6e-05	8.3e-03
GO:0006511	ubiquitin-dependent protein catabolic process	24	2.6e-05	8.3e-03
GO:0006888	ER to Golgi vesicle-mediated transport	24	3.5e-05	1.0e-02
GO:0006886	intracellular protein transport	31	4.3e-05	1.1e-02
GO:1904874	positive regulation of telomerase RNA localization to Cajal body	7	8.3e-05	2.0e-02
GO:0006090	pyruvate metabolic process	8	9.6e-05	2.1e-02
GO:0070125	mitochondrial translational elongation	16	1.1e-04	2.2e-02
GO:0006446	regulation of translational initiation	10	1.5e-04	2.8e-02
GO:0043039	tRNA aminoacylation	5	1.6e-04	3.0e-02
GO:0018107	peptidyl-threonine phosphorylation	10	2.9e-04	4.9e-02
GO:0000462	maturation of SSU-rRNA from tricistronic rRNA transcript	9	3.3e-04	5.4e-02
GO:0006610	ribosomal protein import into nucleus	5	3.7e-04	5.6e-02
GO:0016236	macroautophagy	14	4.0e-04	5.9e-02

RNA-seq: Top age-related GO terms (Cellular Component)

GO term ID	Description	Count	p-value	Benjamini
GO:0016020	membrane	220	9.8e-21	3.7e-18
GO:0005739	mitochondrion	157	1.2e-20	3.7e-18
GO:0070062	extracellular exosome	242	4.3e-16	9.1e-14
GO:0005829	cytosol	282	1.0e-15	1.6e-13
GO:0005913	cell-cell adherens junction	57	9.5e-15	1.2e-12
GO:0005737	cytoplasm	380	2.3e-13	2.4e-11
GO:0043209	myelin sheath	36	4.7e-13	4.2e-11
GO:0005759	mitochondrial matrix	47	5.7e-09	4.5e-07
GO:0005654	nucleoplasm	217	1.1e-08	7.8e-07
GO:0000502	proteasome complex	18	1.4e-08	8.0e-07
GO:0005743	mitochondrial inner membrane	56	1.4e-08	8.0e-07
GO:0042645	mitochondrial nucleoid	14	3.5e-07	1.8e-05
GO:0014704	intercalated disc	14	8.5e-07	4.2e-05
GO:0005832	chaperonin-containing T-complex	7	2.5e-06	1.1e-04
GO:0005643	nuclear pore	16	5.2e-06	2.2e-04
GO:0043231	intracellular membrane-bounded organelle	55	2.7e-05	1.1e-03
GO:0002199	zona pellucida receptor complex	6	2.9e-05	1.1e-03
GO:0043034	costamere	8	5.4e-05	1.9e-03
GO:0043234	protein complex	42	7.8e-05	2.6e-03
GO:0045254	pyruvate dehydrogenase complex	5	1.5e-04	4.6e-03

Conclusion

- GBMs provide a flexible framework for the analysis of matrix data.
- Delta propagation is a novel general technique for uncertainty quantification.
- Our algorithms enable accurate GBM estimation and inference in modern applications.
- Possible directions for future work:
  - extend to more general bilinear model structures,
  - seek theoretical guarantees for delta propagation, and
  - try applying delta propagation to other models.
- Preprint is on arXiv: `https://arxiv.org/abs/2010.04896`

Inference in generalized bilinear models

Jeff Miller

Joint work with Scott L. Carter

Harvard T.H. Chan School of Public Health
Department of Biostatistics

New England Statistics Symposium (NESS)
University of Connecticut, May 24, 2022

Preprint: https://arxiv.org/abs/2010.04896
Slides: http://jwmi.github.io/talks/ness2022.pdf