Learning Latent Representations with Prior Information Using Autoencoders

Sergei Rybakov, TUM

November 21st, 2019

Representation learning and factor models

Current problems with unsupervised representations for single-cell data

- Representations learned by autoencoders are **hard to interpret and explain**.
- Human understanding of data is most effective through decomposition into **interpretable components** based on **prior knowledge.**

Draw from factor modeling ideas

Factor models ”build in” interpretability by regularizing factors with prior knowledge.

Primary objective: Can we use this idea for learning interpretable representations for autoencoders?

Secondary objective: Can representations be learned more efficiently by using prior knowledge?

Prior Knowledge – Pathway Databases

There are many resources for prior knowledge, like MSigDB:

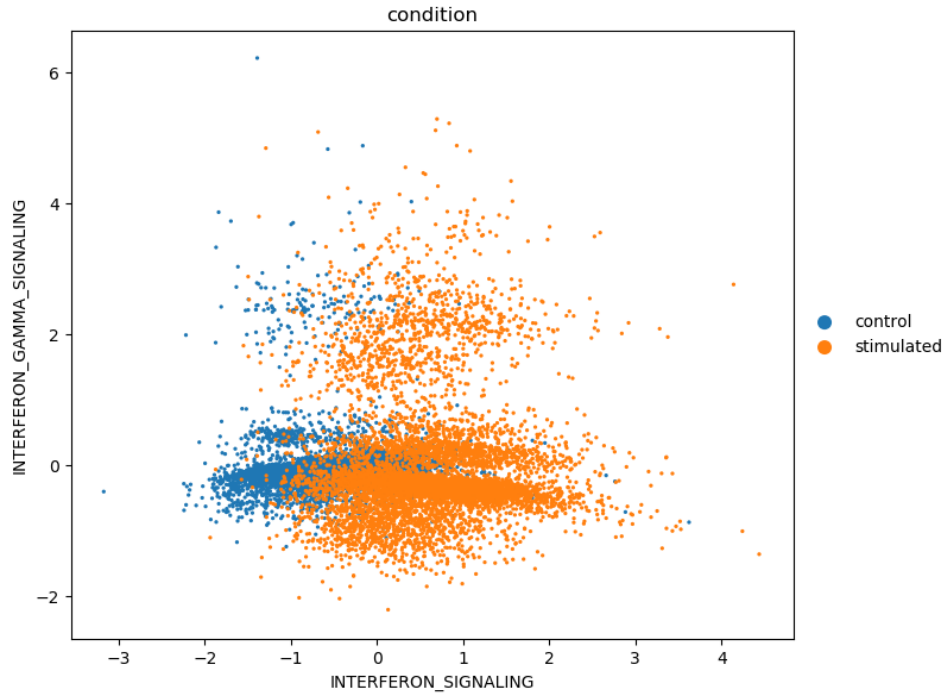
TNFA_SIGNALING_VIA_NFKB http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_TNFA_SIGNALING_VIA_NFKB JUNB CXCL2 ATF3 NFKBIA TNFAIP3 PTGS2
HYPOXIA http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_HYPOXIA PGK1 PDK1 GBE1 PFKL ALDOA ENO2 PGM1 NDRG1 HK2 ALDOC GPI MXI1 SLC
CHOLESTEROL_HOMEOSTASIS http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_CHOLESTEROL_HOMEOSTASIS FDPS CYP51A1 IDI1 FDFT1 DHCR7 SQLE H
MITOTIC_SPINDLE http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_MITOTIC_SPINDLE ARHGEF2 CLASP1 KIF11 AC027237.1 ALS2 ARF6 MYO9B MYH
WNT_BETA_CATENIN_SIGNALING http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_WNT_BETA_CATENIN_SIGNALING MYC CTNNB1 JAG2 NOTCH1 DLL1 A
TGF_BETA_SIGNALING http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_TGF_BETA_SIGNALING TGFBR1 SMAD7 TGFB1 SMURF2 SMURF1 BMPR2 SKIL S
IL6_JAK_STAT3_SIGNALING http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_IL6_JAK_STAT3_SIGNALING IL4R IL6ST STAT1 IL1R1 CSF2RB SOCS3 S
DNA_REPAIR http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_DNA_REPAIR POLR2H POLR2A POLR2G POLR2E POLR2J POLR2F POLR2C POLR2K G
G2M_CHECKPOINT http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_G2M_CHECKPOINT AURKA CCNA2 TOP2A CCNB2 CENPA BIRC5 CDC20 PLK1 TTK PRC1
APOPTOSIS http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_APOPTOSIS CASP3 CASP9 DFFA CASP7 CFLAR BIRC3 PMAIP1 CASP8 JUN BCL2L11 MCL1
NOTCH_SIGNALING http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_NOTCH_SIGNALING JAG1 NOTCH3 NOTCH2 APH1A HES1 CCND1 FZD1 PSEN2 FZD
ADIPOGENESIS http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_ADIPOGENESIS FABP4 ADIPOQ PPARG LIPE DGAT1 LPL CPT2 CD36 GPAM ADIPOR2
ESTROGEN_RESPONSE_EARLY http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_ESTROGEN_RESPONSE_EARLY GREB1 CA12 SLC9A3R1 MYB ANXA9 IGFBP4
ESTROGEN_RESPONSE_LATE http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_ESTROGEN_RESPONSE_LATE TFF1 SLC9A3R1 TPD52L1 PRSS23 CA12 PDZ
ANDROGEN_RESPONSE http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_ANDROGEN_RESPONSE KLK3 KLK2 ACSL3 PIAS1 CAMKK2 NKX3-1 TMPRSS2 APP
MYOGENESIS http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_MYOGENESIS ACTA1 TNNI2 MYL1 TNNC1 TNNC2 MYH3 MYLPF TNNT3 TNNT2 CASQ2 ACTC1
PROTEIN_SECRETION http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_PROTEIN_SECRETION ARCN1 TMED10 COPB2 RAB14 ATP7A COPB1 LAMP2 EGFR I
INTERFERON_ALPHA_RESPONSE http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_INTERFERON_ALPHA_RESPONSE MX1 ISG15 AC004551.1 IFIT3 IFI44 I
INTERFERON_GAMMA_RESPONSE http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_INTERFERON_GAMMA_RESPONSE STAT1 ISG15 IFIT1 MX1 IFIT3 IFI35 I
APICAL_JUNCTION http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_APICAL_JUNCTION ACTN1 CLDN7 ACTN3 CLDN19 DLG1 TJP1 COL17A1 NECTIN1 C
APICAL_SURFACE http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_APICAL_SURFACE B4GALT1 RHCG MAL LYPD3 PKHD1 ATP6V0A4 CRYBG1 SHROOM2 S
HEDGEHOG_SIGNALING http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_HEDGEHOG_SIGNALING SHH PTCH1 NRCAM NRP1 SCG2 AMOT UNC5C ADGRG1 H
COMPLEMENT http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_COMPLEMENT C2 C1S CFB C1R SERPINE1 MMP14 SERPING1 CTSL F5 MMP13 F7 CTS
UNFOLDED_PROTEIN_RESPONSE http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_UNFOLDED_PROTEIN_RESPONSE ATF4 HERPUD1 PARN EXOSC4 HSP90B1
PI3K_AKT_MTOR_SIGNALING http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_PI3K_AKT_MTOR_SIGNALING MAPK8 PIK3R3 GRB2 NFKBIB MAP2K6 MAP
MTORC1_SIGNALING http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_MTORC1_SIGNALING FADS1 DDIT4 CALR HK2 PGK1 SLC7A5 CTSC ACSL3 SLC1A
E2F_TARGETS http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_E2F_TARGETS AURKA BRCA2 CCP110 CENPE CKS2 DCLRE1B DNMT1 DONSON EED GINS1
MYC_TARGETS_V1 http://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_MYC_TARGETS_V1 PCNA PSMD8 PSMD7 SET SNRPA1 RAN SRSF2 G3BP1 STARD7 NPM

Application examples (1):

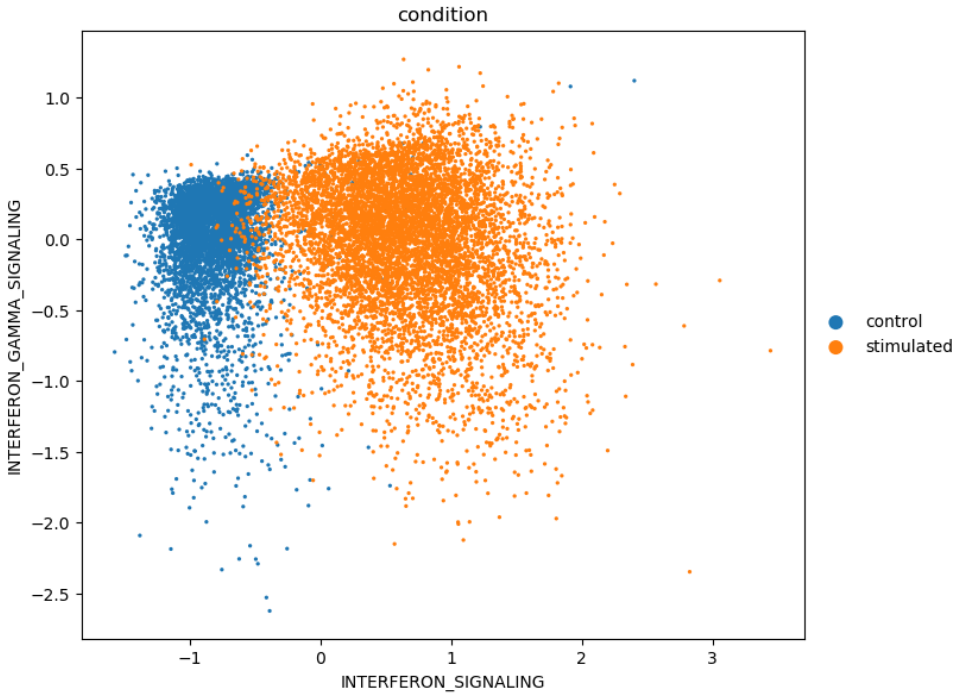
Infer pathway activation in immune response stimulation (Kang17)

Interferon gamma pathway, immune-related pathways should be active for condition=stimulated.

f-scLVM

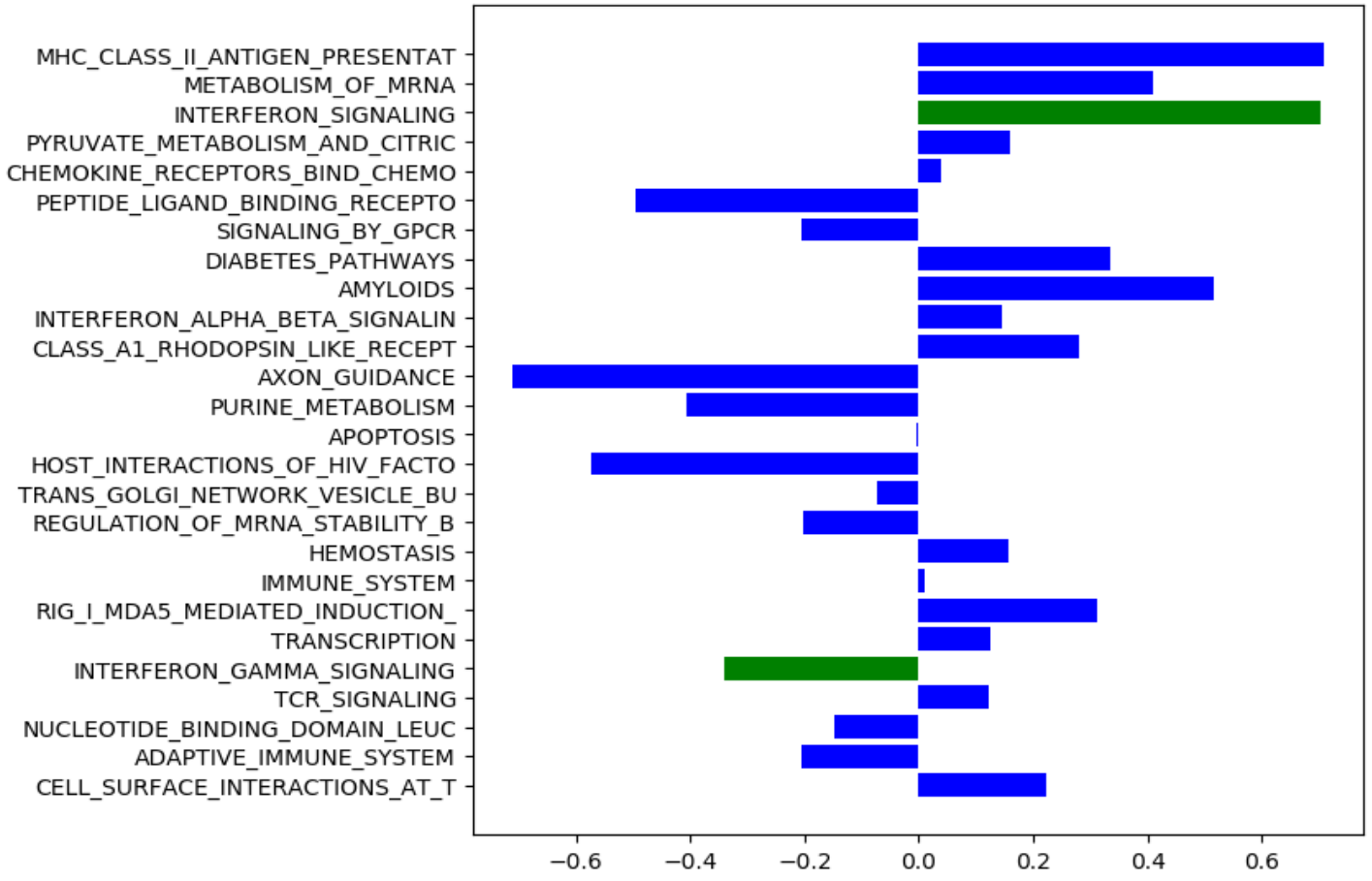


Interpretable Autoencoder



Application examples (1):

Correlations between loadings in f-scLVM and Interpretable Autoencoder
for top relevance factors (relevance from f-scLVM).
Green – interferon signaling, interferon gamma signaling

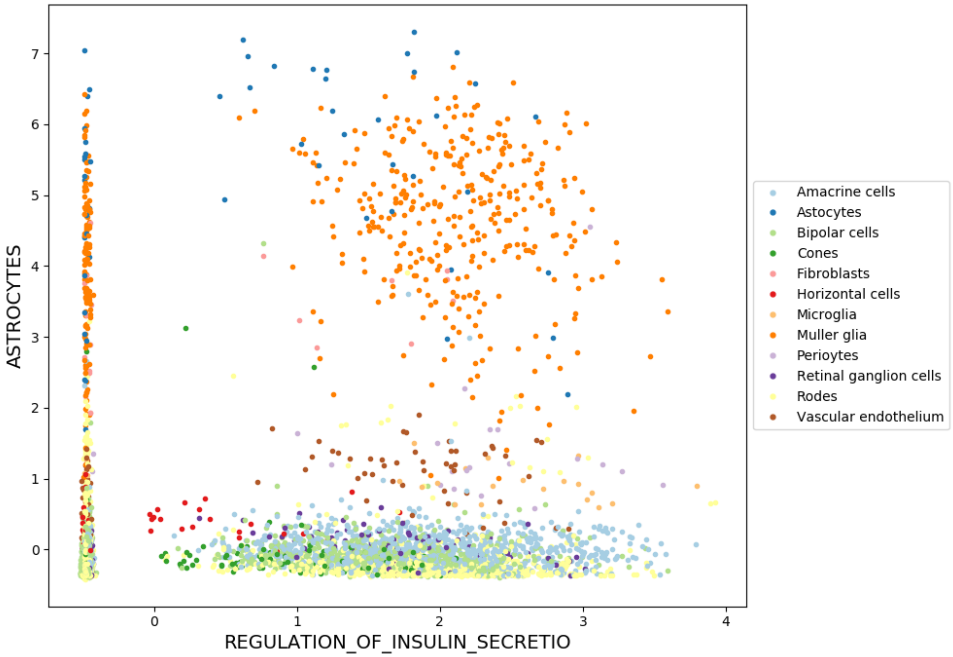


Application examples (2)

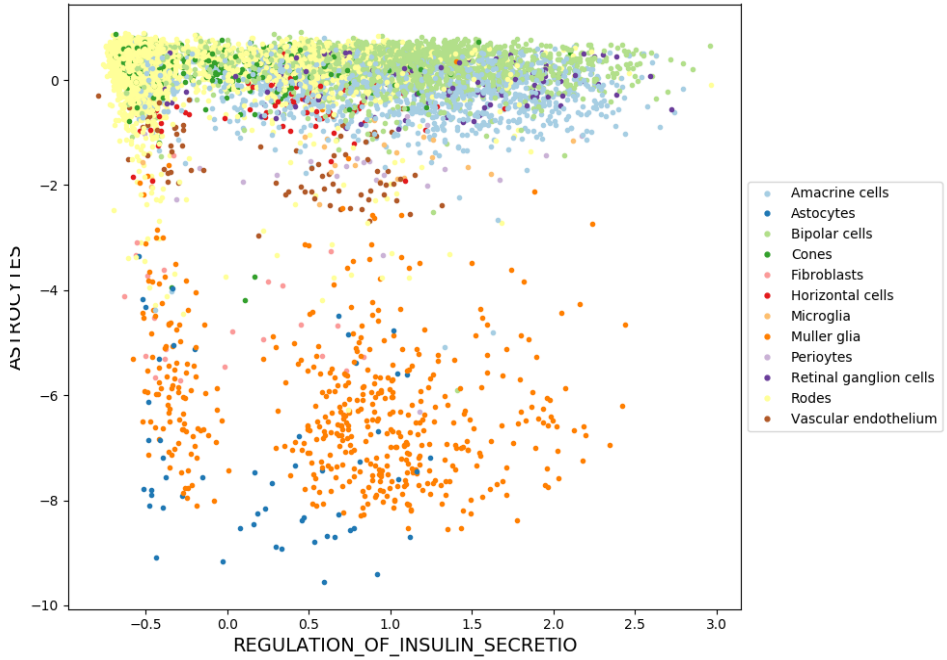
Infer cell identity jointly with pathway activation (Macosko15)

Infer insulin stimulation of astrocytes for a subset of 50k retina cells.

f-scLVM



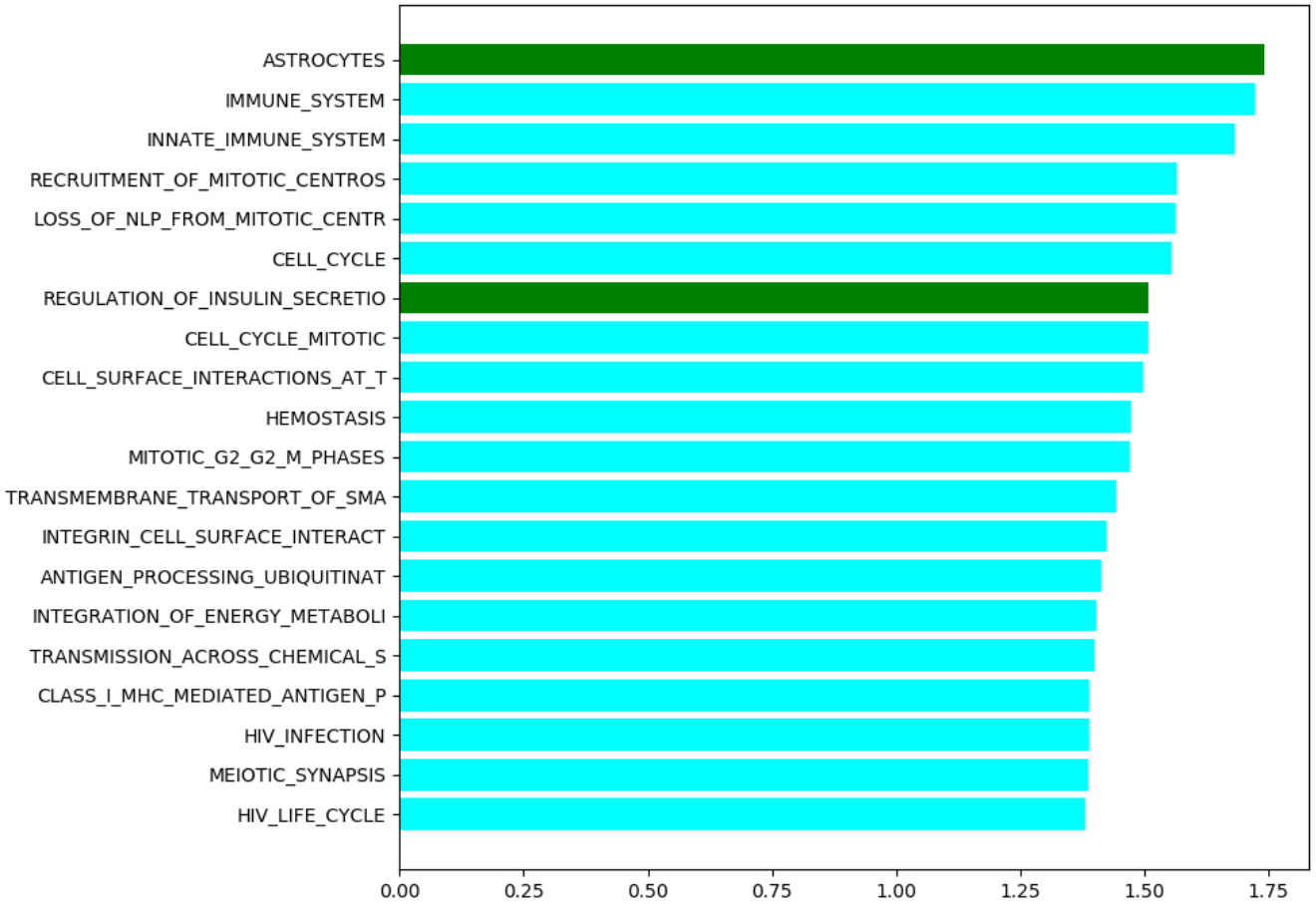
Interpretable Autoencoder



Application examples (2):

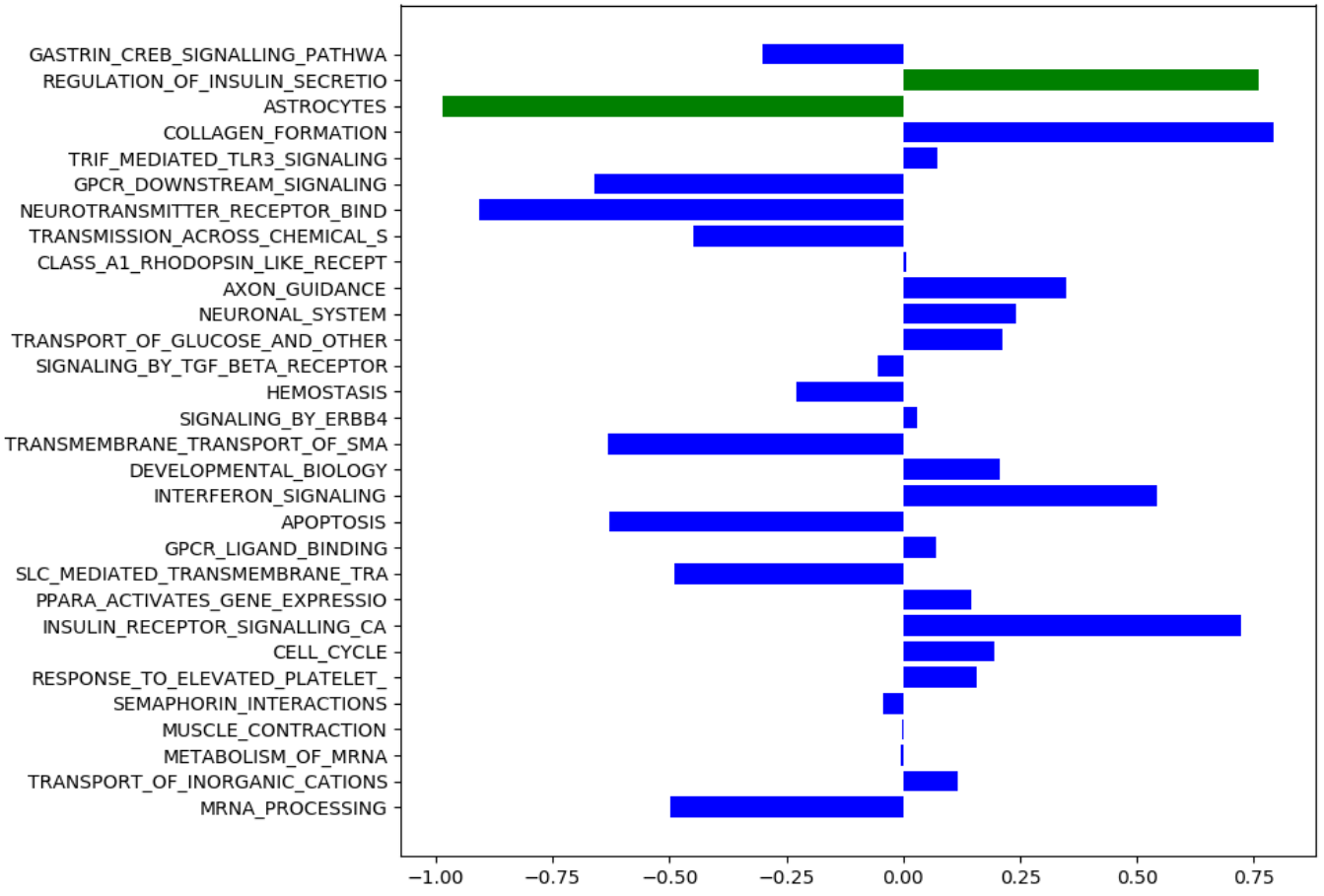
Top terms by weights' norm in Interpretable Autoencoder

Green – Astrocytes, regulation of insulin secretion



Application examples (2):

Correlations between loadings in f-scLVM and Interpretable Autoencoder
for top relevance factors (relevance from f-scLVM).
Green – Astrocytes, regulation of insulin secretion



Runtime comparison

Orders of magnitude difference in runtime:

For a dataset of size ~ 10 000 x 1000 with ~ 140 annotated terms on ICB servers:

f-scLVM (Slalom python package) ~ **2 full days**

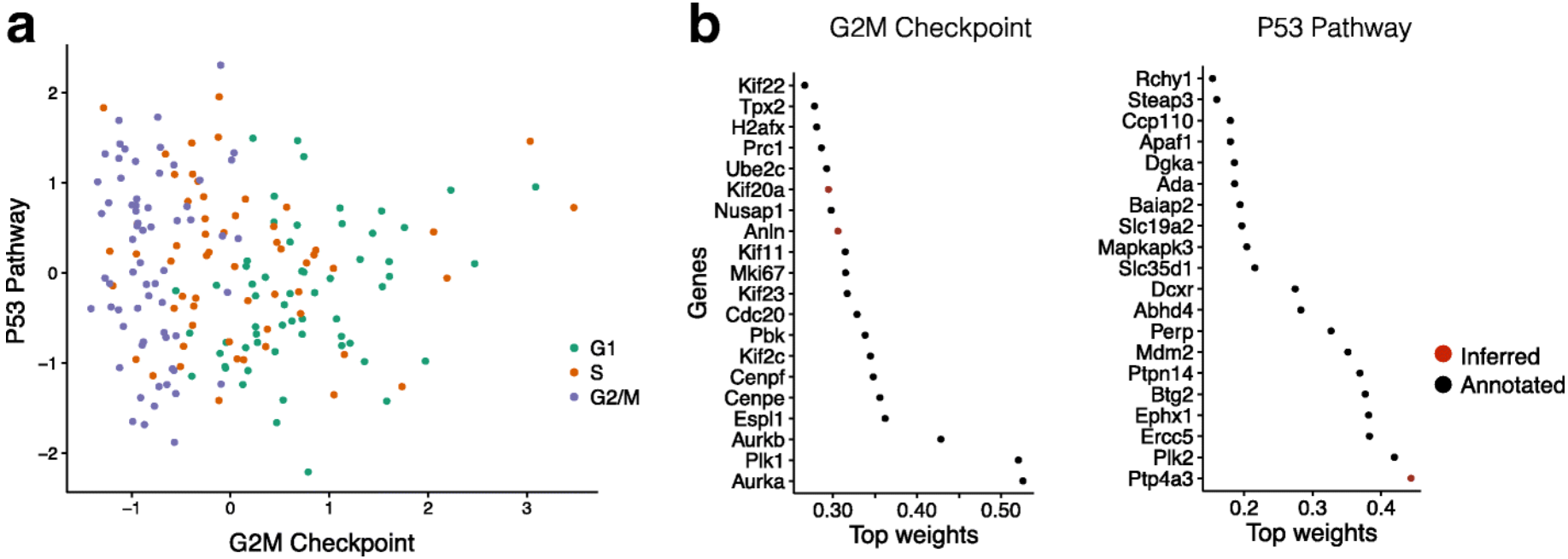
Autoencoder (80 epochs, CPU) ~ 30minutes
Can be improved, accelerated by GPUs.

Application examples (3)

Correcting prior knowledge (Buettner15, Buettner18)

Application of f-scLVM to 182 (!) mouse embryonic stem cells, experimentally staged for the cell cycle.

f-scLVM meaningfully corrects gene sets?
182 cells not enough for interpretable autoencoder.

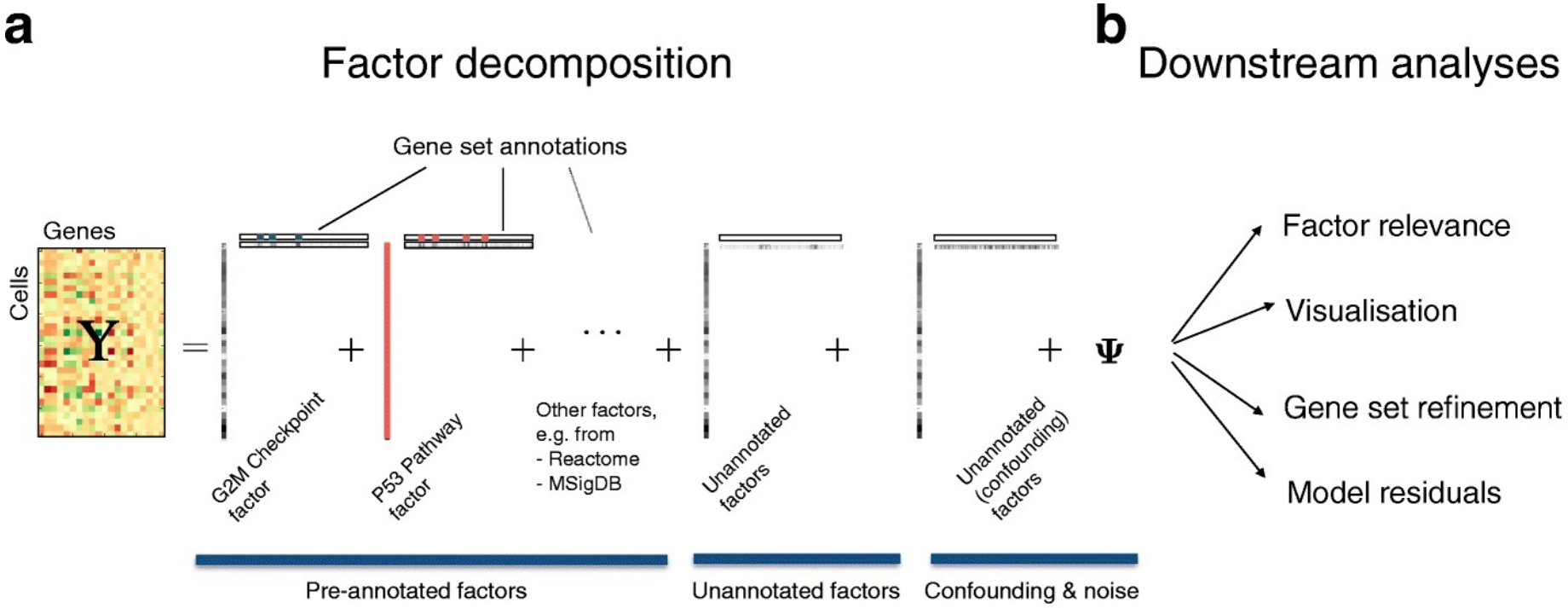


f-scLVM (2): Outline

Factorial single-cell latent variable model (f-scLVM) is a Bayesian model.

f-scLVM is based on a variant of factor analysis.

It decomposes the matrix of single-cell gene expression profiles into factors and weights.



f-scLVM (2): Basic equations

$$Y = \sum_{c=1}^{C} u_c V_c^T + \sum_{a=1}^{A} p_a R_a^T + \sum_{h=1}^{H} s_h Q_h^T + \Psi$$

Or in the matrix form: $$Y = XW^T + \Psi$$

$$X = [u_1, \ldots, u_C, p_1, \ldots, p_A, s_1, \ldots, s_H]$$

$$W = [V_1, \ldots, V_C, R_1, \ldots, R_A, Q_1, \ldots, Q_H]$$

Y denotes the gene expression matrix where rows correspond to each of *N* cells and columns correspond to G genes.

The column vectors u_c, p_a, s_h represent the known cell covariates, as well as cell states for annotated and unannotated factors

The column vectors V_c, R_a, Q_h are the corresponding regulatory weights of a given factor on all genes.

The matrix Ψ denotes Gaussian residual noise. $$\Psi \sim \mathcal{N}(0, \mathrm{diag}(\tau^{-1}))$$

f-scLVM (3): the first level of regularization

Two levels of regularization on the corresponding columns of the weight matrix W are employed.

The first level is a gene-level sparsity prior on the elements of individual columns of W.

$$P(W|Z) = \prod_{g=1}^{G} \prod_{k=1}^{K} P(w_{g,k} \mid z_{g,k})$$

$$P(w_{g,k} \mid z_{g,k}) = \begin{cases} \mathcal{N}(w_{g,k} \mid 0, \frac{1}{\alpha_k}), & \text{if } z_{g,k} = 1 \\ \delta_0(w_{g,k}), & \text{if } z_{g,k} = 0 \end{cases}$$

$$P(I^n_{g,k} \mid z_{g,k}) = \begin{cases} \mathrm{Ber}(I^n_{g,k} \mid 1 - \mathrm{FPR}), & \text{if } z_{g,k} = 1 \\ \mathrm{Ber}(I^n_{g,k} \mid \mathrm{FNR}), & \text{if } z_{g,k} = 0 \end{cases}$$

The binary variable $z_{g,k}$ determines whether factor k has as a regulatory effect on gene or not.

The true state of the indicator variable $z_{g,k}$ is unobserved; however, for annotated factors the pathway annotations provide partial evidence

FNR=0.001 and FPR=0.01

f-scLVM (4): the second level of regularization

The second level of regularization is an automatic relevance determination prior on the factor level which **deactivates the factors that are unused**.

$$P(\alpha_k) = \text{Gamma}(\alpha_k \mid a_\alpha, b_\alpha)$$

For factors that do not explain variation in the data the precision α_k will be large,

Also

$$P(Y \mid X, W, \tau) = \prod_{n=1}^{N} \mathcal{N}(y_n \mid x_n W^T, \text{diag}(\tau^{-1}))$$

Where $\text{diag}(\tau^{-1})$ denotes the diagonal covariance matrix formed of the inverse elements of the noise precisions for each dimension (gene) $\tau = (\tau_1, \ldots, \tau_G)$.

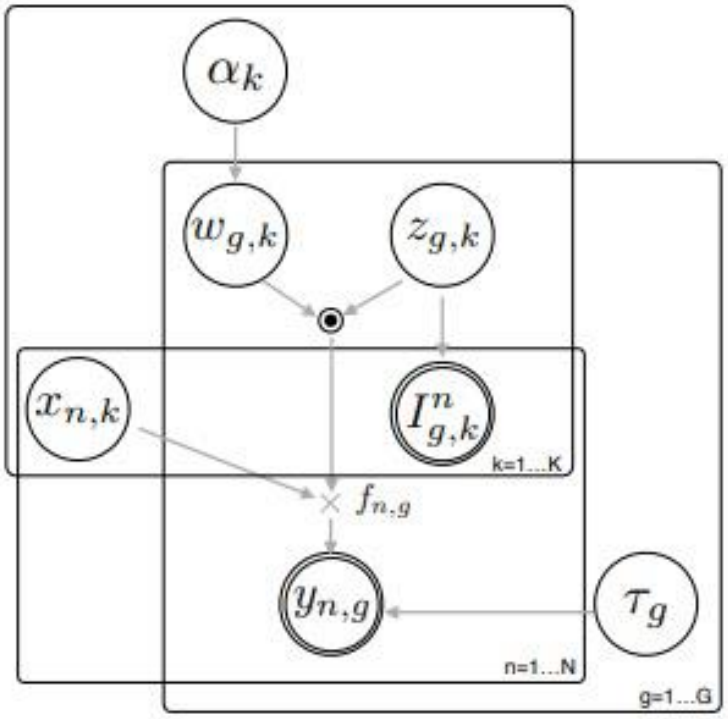
$$P(Y, I, X, W, Z, \tau) = \prod_n \mathcal{N}(y_n \mid x_n W^T, \text{diag}(\tau^{-1})) \cdot \prod_{n,k} \mathcal{N}(x_{n,k} \mid 0, 1) \cdot \prod_{n,g,k} P(I^n_{g,k} \mid z_{g,k}) \cdot \cdot \prod_{g,k} P(w_{g,k} \mid z_{g,k}) P(z_{g,k}) \cdot \prod_g P(\tau_g) \cdot \prod_k P(\alpha_k)$$

f-scLVM (5): Graphical model form

The circled variables are random and unobserved.

The double-circled variables denote observed data.

Statistical dependencies between all variables are indicated using arrows.



Why not enrichment tools?

Enrichment tools score each observation for activation of gene sets independent from all other gene sets.

This univariate treatment typically leads to many false positives.

In addition, the commonly used tools like GSEA, show very long runtimes.

For single-cell: Cell set enrichment analysis (CSEA)

Schafflick19

https://www.biorxiv.org/content/10.1101/403527v2

Undercomplete autoencoders

Definition:

$$\text{input } Y \in \mathbb{R}^{N \times G}$$
$$Z = f_{W_1}(Y) \in \mathbb{R}^{N \times H}$$
$$g_{W_2}(Z) \in \mathbb{R}^{N \times G}$$
$$Y \approx g_{W_2}(f_{W_1}(Y))$$

Objective function:

$$L(W_1, W_2) = \sum_{n=1}^{N} ||y_n - g_{W_2}(f_{W_1}(y_n))||_2^2$$

What doesn't work

Hard-coded weight masks in the first layer of an autoencoder

$$y(W \odot I)^T$$

Stochastic weight masks

$$y(W \odot M)^T$$

$$M \sim \mathrm{Ber}(\theta)$$

$$M \approx I$$

Autoencoder with regularized linear decoder (1): Outline

Nonlinear encoder f

Linear decoder g

$$Y \approx g_W(f_\theta(Y)) = f_\theta(Y)W^T$$

Idea – regularize decoder with prior knowledge

$$L(\theta, W) = \frac{1}{N}\sum_{n=1}^{N} ||y_n - f_\theta(y_n)W^T||_2^2 + \frac{\lambda_0}{N}\sum_{n=1}^{N} ||f_\theta(y_n)||_2^2 + R_{\lambda_1,\lambda_2,\lambda_3}(W)$$

R – structured sparsity enforcing regularization for W

Autoencoder with regularized linear decoder (2): Regularization

Two levels of regularization as in f-scLVM.

$$R_{\lambda_1,\lambda_2,\lambda_3}(W) = R^1_{\lambda_1,\lambda_2}(W) + \lambda_3 R^2(W)$$

First level – enforce sparsity for inactive genes.

$$R^1_{\lambda_1,\lambda_2}(W) = \lambda_1 \sum_{k_1} ||W_{:,k} \odot (1 - I_{:,k})||_1 + \lambda_2 \sum_{k_2} ||W_{:,k}||_1$$

Second level – enforce sparsity on the level of factors (deactivate irrelevant factors).

$$R^2(W) = \sum_k ||W_{:,k}||_2$$

Gathering all terms

$$R_{\lambda_1,\lambda_2,\lambda_3}(W) = \lambda_1 \sum_{k_1} ||W_{:,k} \odot (1 - I_{:,k})||_1 + \\ + \lambda_2 \sum_{k_2} ||W_{:,k}||_1 + \lambda_3 \sum_k ||W_{:,k}||_2$$

Autoencoder with regularized linear decoder (3): link to f-scLVM

The form of the regularization function is motivated by the **negative logarithm of the posterior distribution** of f-scLVM.

$$-\log P(Y, I, X, \widetilde{W}, Z, \tau) = \sum_n ||y_n - x_n(\widetilde{W} \odot Z)^T||_2^2 + \\ + \sum_{n,g,k} -\log P(I_{g,k}^n \mid z_{g,k}) + \sum_{g,k} -\log P(z_{g,k}) + \dots$$

Negative logarithms of the priors for Z are equivalent to

$$R^1(W) = \alpha_2 \sum_{k_2} ||W_{:,k}||_0 + \alpha_3 \sum_{k_3} ||W_{:,k}||_0 + \\ + \alpha_4 \sum_{k_1} ||W_{:,k} \odot I_{:,k}||_0 + \alpha_5 \sum_{k_1} ||W_{:,k} \odot (1 - I_{:,k})||_0$$

Coefficient for dense unannotated factors (a3) and active genes in annotated factors (a4) are **negative.**

Replace L0 “norm” with L1 norm, omit the terms with the negative coefficients

Training with stochastic proximal gradient descent

Non-differentiable points in regularization makes stochastic gradient descent intractable.

Proximal operators allow to circumvent this and enforce sparsity efficiently.

$$F(\theta, W) = \frac{1}{N}\sum_{n=1}^{N} ||y_n - f_\theta(y_n)W^T||_2^2 + \frac{\lambda_0}{N}\sum_{n=1}^{N} ||f_\theta(y_n)||_2^2$$

Update scheme

$$\theta^{(t+1)} = \theta^{(t)} - \eta\nabla_\theta \hat{F}(\theta, W)$$
$$W^{(t+1)} = \underset{\eta R_{\lambda_1,\lambda_2,\lambda_3}}{\text{Prox}}(W^{(t)} - \eta\nabla_W \hat{F}(\theta, W))$$

Proximal operator

$$\underset{\eta R_{\lambda_1,\lambda_2,\lambda_3}}{\text{Prox}}(V) = \arg\min_L \frac{1}{2}||L - V||_F^2 + \eta R_{\lambda_1,\lambda_2,\lambda_3}(L)$$

Proximal operators (1)

$$R_{\lambda_1,\lambda_2,\lambda_3}(W) = \lambda_1 \sum_{k_1} ||W_{:,k} \odot (1 - I_{:,k})||_1 + \\ + \lambda_2 \sum_{k_2} ||W_{:,k}||_1 + \lambda_3 \sum_{k} ||W_{:,k}||_2$$

The regularization summand for a separate column k of W can be written as one of the three forms below

$$R^{k_1,k}_{\lambda_1,\lambda_3}(W_{:,k}) = \lambda_1 ||W_{:,k} \odot (1 - I_{:,k})||_1 + \lambda_3 ||W_{:,k}||_2$$
$$R^{k_2,k}_{\lambda_2,\lambda_3}(W_{:,k}) = \lambda_2 ||W_{:,k}||_1 + \lambda_3 ||W_{:,k}||_2$$
$$R^{k_3,k}_{\lambda_3}(W_{:,k}) = \lambda_3 ||W_{:,k}||_2$$

Where the specific form depends on the membership of the factor column k in the set of annotated factors (k_1), sparse unannotated factors (k_2) or dense unannotated factors (k_3).

Proximal operators (2)

$$\underset{\eta R^{k_3,k}_{\lambda_3}}{\text{Prox}}(v) = \begin{cases} v - \eta\lambda_3 \frac{v}{||v||_2}, & \text{if } ||v||_2 > \eta\lambda_3 \\ 0, & \text{if } ||v||_2 \leq \eta\lambda_3 \end{cases}$$

$$\underset{\eta R^{k_2,k}_{\lambda_2,\lambda_3}}{\text{Prox}}(v) = \underset{\eta\lambda_3||\cdot||_2}{\text{Prox}}\left(\underset{\eta\lambda_2||\cdot||_1}{\text{Prox}}(v)\right)$$

$$\mathcal{T}_{\lambda_2}(y) = \begin{cases} y - \lambda_2, & \text{if } y \geq \lambda_2 \\ 0, & \text{if } |y| < \lambda_2 \\ y - \lambda_2, & \text{if } y \leq \lambda_2 \end{cases}$$

$$\underset{\lambda_2||\cdot||_1}{\text{Prox}}(v) = \mathcal{T}_{\lambda_2}(v_1) \times \mathcal{T}_{\lambda_2}(v_2) \times \cdots \times \mathcal{T}_{\lambda_2}(v_G)$$

$$\underset{\eta R^{k_1,k}_{\lambda_1,\lambda_3}}{\text{Prox}}(v) = \underset{\eta\lambda_3||\cdot||_2}{\text{Prox}}\left(\underset{\eta\lambda_1||\cdot\odot(1-I_{:,k})||_1}{\text{Prox}}(v)\right)$$

$$\mathcal{A}^{g,k}_{\lambda_1}(y) = \begin{cases} y, & \text{if } I_{g,k} = 1 \\ \mathcal{T}_{\lambda_1}(y), & \text{if } I_{g,k} = 0 \end{cases}$$

$$\underset{\lambda_1||\cdot\odot(1-I_{:,k})||_1}{\text{Prox}}(v) = \mathcal{A}^{1,k}_{\lambda_1}(v_1) \times \mathcal{A}^{2,k}_{\lambda_1}(v_2) \times \cdots \times \mathcal{A}^{G,k}_{\lambda_1}(v_G)$$

Motivation for the omission of the terms with negative coefficients

Proximal operator for L0 "norm" with a negative coefficient

operator $\underset{-||\cdot||_0}{\text{Prox}}(v)$ can be written as

$$q_y(z) = \begin{cases} (y-z)^2 - 1, & \text{if } z \neq 0 \\ y^2, & \text{if } z = 0 \end{cases}$$

$$\underset{-||\cdot||_0}{\text{Prox}}(v) = \arg\min_z q_{v_1}(z) \times \arg\min_z q_{v_2}(z) \times \cdots \times \arg\min_z q_{v_G}(z)$$

However, it can be clearly seen that

$$\arg\min_z q_y(z) = \begin{cases} y, & \text{if } y \neq 0 \\ \varnothing, & \text{if } y = 0 \end{cases}$$

The proximal operator reduces to the identity function because solutions won't be sparse.

Summary: f-scLVM vs Autoencoder

f-scLVM

Pro:
- Can be used with (very) small datasets.
- (almost) No hyperparameter tuning.

Contra:
- Inefficient implementation.
- No inference for out-of-sample data.

Interpretable autoencoder

Pro:
- **Much better** scalability for large datasets.
- Can use GPUs.
- Inference for out-of-sample data.

Contra:
- Requires large datasets to train it.
- **Manual hyperparameter tuning.**