Title
Dictionary learning: analysis of spatial gene expression data and local identifiability theory

Author
Wu, Siqi

Publication Date
2016

Peer reviewed|Thesis/dissertation

Dictionary learning: analysis of spatial gene expression data and local identifiability theory

by

Siqi Wu

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bin Yu, Chair
Professor Peter Bickel
Professor Bruno Olshausen

Spring 2016

Dictionary learning: analysis of spatial gene expression data and local identifiability theory

Abstract

Dictionary learning: analysis of spatial gene expression data and local identifiability theory

by

Siqi Wu

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Bin Yu, Chair

Spatial gene expression data enable the detection of local covariability and are extremely useful for identifying local gene interactions during normal development. The abundance of spatial expression data in recent years has led to the modeling and analysis of regulatory networks. The inherent complexity of such data makes it a challenge to extract biological information. In the first part of the thesis, we developed staNMF, a method that combines a dictionary learning algorithm called nonnegative matrix factorization (NMF), with a new stability-driven criterion to select the number of dictionary atoms. When applied to a set of *Drosophila* early embryonic spatial gene expression images, one of the largest datasets of its kind, staNMF identified a dictionary with 21 atoms, which we call *principal patterns* (PP). Providing a compact yet biologically interpretable representation of *Drosophila* expression patterns, PP are comparable to a fate map generated experimentally by laser ablation and show exceptional promise as a data-driven alternative to manual annotations. Our analysis mapped genes to cell-fate programs and assigned putative biological roles to uncharacterized genes. Furthermore, we used the PP to generate local transcription factor (TF) regulatory networks. Spatially local correlation networks (SLCN) were constructed for six PP that span along the embryonic anterior-posterior axis. Using a two-tail 5% cut-off on correlation, we reproduced 10 of the 11 links in the well-studied gap gene network. The performance of PP with the *Drosophila* data suggests that staNMF provides informative decompositions and constitutes a useful computational lens through which to extract biological insight from complex and often noisy gene expression data.

The biological interpretability of the NMF-derived dictionary motivated us to understand why dictionary learning works analytically. In particular, if the observed data are generated from a ground truth dictionary, under what conditions can dictionary learning recovers the true dictionary? In the second part of the thesis, we studied the local correctness, or *local identifiability*, of a particular dictionary learning formulation with the l_1-norm objective function. Suppose we observe N data points $\mathbf{x}_i \in \mathbb{R}^K$ for $i = 1, ..., N$, where $\mathbf{x}_i$'s are $i.i.d.$ random linear combinations of the K columns from a square and invertible dictionary $\mathbf{D}_0 \in \mathbb{R}^{K \times K}$. We assumed that the random linear coefficients are generated from either

the s-sparse Gaussian model or the Bernoulli-Gaussian model. For the population case, we established a sufficient and almost necessary condition for $\mathbf{D}_0$ to be locally identifiable, i.e., a local minimum of the expected l_1-norm objective function. Our condition covers both sparse and dense cases of the random linear coefficients and significantly improves the sufficient condition in Gribonval and Schnass (2010). Moreover, we demonstrated that for a complete μ-coherent reference dictionary, i.e., a dictionary with absolute pairwise column inner-product at most $\mu \in [0, 1)$, local identifiability holds even when the random linear coefficient vector has up to $O(\mu^{-2})$ nonzeros on average. Finally, it was shown that our local identifiability results translate to the finite sample case with high probability provided $N = O(K \log K)$.

Contents

List of Figures

List of Tables

Acknowledgments

First of all, I would like to thank my advisors Professor Bin Yu and Dr. Erwin Frise. Prof. Yu is known to have a strict mentoring style and high expectations for her students. Working with her, I benefited not only from her insightful ideas on applied and theoretical statistics, but also from her high standard for conducting scientific research. In addition to her academic assistance, Prof. Yu showed intense care for my career and personal life. She always encourages me to live a meaningful life and to give back to the society that has made my education possible. She is definitely a wonderful role model for me to follow. Dr. Frise, of Lawrence Berkeley National Laboratory, graciously offered me tremendous support on the biology and computing ends. He introduced me to the exciting field of systems biology, and has always been patient in explaining to me fundamentals of biology. Working with Prof. Yu and Dr. Frise has been such a precious experience that I will treasure for the rest of my life.

I am grateful to all members of the Yu Group. In particular, I would like to thank visiting scholar Xiangyu Chang for first introducing me to the group and encouraging me to work with Prof. Yu; Julien Mairal for the very helpful discussions and all the technical supports for using his dictionary learning package; Yuansi Chen, Antony Joseph, Karl Kumbier and Yu Wang for working with me closely on a number of applied and theoretical projects; Hongwei Li and Sivaraman Balakrishnan for giving me extremely useful career guidance. I would also like to thank group members Reza Abbasi-Asl, Rebecca Barter, Yuval Benjamini, Adam Bloniarz, Sumanta Basu, Raaz Dwivedi, Jingxue Fu, Christine Kuang, Hanzhong Liu, Xiusheng Lu, Taesup Moon, Jamie Murdoch, Sujayam Saha, Simon Walter and Shijing Yao for their constant support.

I am indebted to members of the Berkeley *Drosophila* Genome Project (BDGP). I would like to thank Dr. Ben Brown for connecting Dr. Frise with Prof. Yu and introducing me to this fantastic project that ultimately became a part of my thesis; Dr. Susan Celniker and Dr. Ann Hammonds for providing valuable biological insights into our statistical analysis and performing large-scale experiments to validate our predictions; Dr. William Fisher for coaching me on how to use sophisticated microscopes. I have been incredibly fortunate to work with this team of world-renowned expert biologists.

I appreciate the considerable help from faculty members and friends at Berkeley. In particular, I would like to thank Professors Ani Adhikari, Peter Bartlett, Peter Bickel, Steven Evans, Adityanand Guntuboyina, Haiyan Huang, Michael Klass, Jim Pitman and Martin Wainwright; my fellow students Riddhipratim Basu, Xinyan Chen, Hye Soo Choi, Lihua Lei, Suzette Puente, Geoffery Schiebinger, Funan Shi, Wenpin Tang, Linda Tran, Rachel Wang, Yuting Wei, Jason Wu, Fanny Yang, Yumeng Zhang and Angie Zhu.

Citadel LLC generously offered me the Citadel Fellowship to cover one year of my tuition and living expenses at Berkeley. Both projects presented in this thesis were conducted throughout the duration of the fellowship. Citadel's support has greatly alleviated my financial burden and allowed me to concentrate on these projects.

I would also like to take this opportunity to thank the professors of my undergraduate study: Professors Stephen Lee, Wai Keung Li and Hailiang Yang of the University of Hong Kong; Prof. Kung-Sik Chan of the University of Iowa and; Prof. George Roussas of UC Davis – for their time and patience in supervising me in a number of projects. Without their encouragement and support, it would have been impossible for me to pursue my degree at Berkeley.

Lastly, I would like to express my gratitude towards my family. The financial and emotional support from my parents has been vital for my education at Berkeley. I thank my grandmother, aunt and uncle, brother and sister, for their care and constant encouragement. I would also like to thank my spiritual family – all the saints at the Church in Berkeley for their unceasing shepherding and remembering of me in their prayers.

Chapter 1

Introduction

Biological processes in multicellular organisms depend on spatial and temporal control of gene expression. Gene products function in the context of other spatially localized gene products and these interactions have been well characterized for development and tissue differentiation. Recent studies of prenatal [1] and adult human brain [2] revealed widespread anatomical variability in gene networks, which is reflective of developmental processes and of the distribution of major cell types. Spatially resolved studies of tumors uncovered widespread intra-tumor heterogeneity [3, 4, 5, 6, 7, 8]. Given the importance of spatio-temporal gene expression, many efforts are underway to characterize it genome wide. Systematic datasets include *Drosophila* gene expression during embryogenesis (Berkeley *Drosophila* Genome Project (BDGP), [9]) and oogenesis [10], subcellular mRNA localization [11], and in brain [12], imaginal discs [13], central nervous system [14] and other developmental model systems (e.g. *Xenopus* [15], *Ciona* [16] and mouse [17, 18, 19]).

Spatial datasets are complex and quickly surpass the human ability to interpret them. To represent, search and analyze such large spatial expression datasets, they are commonly curated with defined controlled vocabulary [9, 17, 18, 19, 20, 21]. Curation using ontologies is time consuming and requires expert knowledge. Despite significant progress towards automatic computer annotation through supervised learning based on human labels [22, 23, 24, 25, 26], the subtleties inherent in spatial expression patterns are difficult to capture and finding related patterns is challenging. An alternative, complementary to ontologies, is the spatial expression information extracted directly from images [12, 17, 18, 19, 22, 27, 28, 29, 30]. We discovered putative gene interactions by correlating gene expression and performing cluster analysis [27] and others have used sparse Gaussian graphical models [30] to do the same. Due to data complexity and the large size of image collections, image based approaches are not routinely used for modeling.

Organ systems develop through the combinatorial action of gene regulatory networks [21, 31], and gene function and regulatory interactions can markedly differ depending on the spatial location [32]. Studies of genomic enhancer elements have shown that wild-type spatial expression patterns are actually the product of multiple genomic elements. These previous studies dissected biological enhancers and discovered that complex expression patterns could

be subdivided into smaller regions [33, 34]. In *Drosophila*, clustering early embryonic gene expression patterns recovered groups of cells that likely interact with one another, contributing to the formation of organs and tissues [27, 33]. These regions are similar to those identified in studies using laser ablation to determine cell lineage and function [35, 36]. Yakoby et al. proposed an innovative method to model spatial gene expression in *Drosophila* follicle cells as a Boolean combination of smaller building blocks [10]. Due to the small number of gene expression patterns in their work (81 genes), they were able to produce building blocks manually. Such an approach is intuitive and conceptually supported by the aforementioned works on genomic enhancers.

1.1 Dictionary learning to analyze spatial gene expression patterns

In the first part of the thesis (Chapter 2 – 6), we describe a method that interprets, represents and analyzes comprehensive spatial gene expression datasets. Specifically, we adapted a powerful dictionary learning algorithm, nonnegative matrix factorization (NMF) [37], to learn data-driven representations from large and complex data. Given a data set, dictionary learning derives a matrix, called *dictionary*, such that each data point can be expressed as a linear combination of the columns (or atoms) of the dictionary. Constraints and/or penalties are often imposed on the dictionary matrix and the linear coefficients. For example, in the seminal paper of [38], a sparsity penalty was imposed on the linear coefficients to obtain a sparse code from natural scene image batches that corresponds to a family of receptive fields similar to those found in primary visual cortex. NMF, on the other hand, requires both the dictionary and the coefficients to be nonnegative. These constraints enable NMF to learn "parts-based" representations of objects [37]. NMF has been applied to many fields such as image processing and computer vision [39], text mining [40], audio signal separation [41] and bioinformatics [42]. See [43] for a recent review of NMF.

NMF depends on a single parameter, the number K of dictionary atoms. Choosing this parameter has been a challenging task [42, 44, 45]. In Chapter 3, following the stability principle [46], we proposed a new stability-based NMF model selection criterion. Specifically, we exploited the fact that the NMF implementation uses alternating minimization [47], and reasoned that a good NMF-generated dictionary would be stable when perturbing the initializations. For each K, we repeat NMF multiple times with randomly sampled data points as initial inputs and quantify the instability of the resulting dictionaries using an Amari type measure. We select K that achieves the lowest instability of the dictionaries. We called this procedure *stability-driven NMF*, or *staNMF*. The validity of staNMF was confirmed by a number of synthetic datasets generated from known dictionaries.

We applied staNMF to a dataset of 1640 spatial gene expression images during early *Drosophila* embryogenesis and identified 21 dictionary atoms. These dictionary atoms are comparable to regions in the pre-organ fate map mentioned earlier [35, 36]. We called these

dictionary atoms *principal patterns (PP)*, as they can be interpreted as spatial building blocks of the *Drosophila* embryo. Using the sparse model selection method LASSO [48] followed by nonnegative least squares (NLS), we represented each gene expression pattern as a sparse and nonnegative linear combination of the 21 PP. We used this representation to predict annotation labels from a control vocabulary [9] and showed that the PP can serve as an alternative to the traditional annotation approach.

Using the PP-based linear representation for expression patterns, we grouped genes into overlapping categories and assigned putative biological roles to previously uncharacterized genes. To understand gene-gene interaction based on expression patterns, we built spatially local correlation networks (SLCN) to relate transcription factors (TF) in six spatial regions that span along the embryonic anterior-posterior axis. The constructed networks correctly reproduced 10 of 11 links in the well-studied gap gene network [49, 50, 51]. Our approach has the significant potential to become the standard lens for spatial gene expression patterns and play a critical role in advancing the discovery and modeling of spatially localized gene networks.

1.2 Understanding dictionary learning: a sufficient and almost necessary condition for local identifiability

The biological interpretability of the NMF-derived dictionary motivated us to understand the theoretical properties of dictionary learning. Despite the empirical success of many dictionary learning formulations [38, 37, 47], relatively little theory is available to explain why they work. One line of research addresses the problem of *dictionary identifiability* [52, 53, 54]: if the data vectors are generated as sparse linear combinations of the atoms of a true dictionary $\mathbf{D}_0$, under what conditions can we recover $\mathbf{D}_0$ by solving the dictionary learning problem? In the second part of the thesis (Chapter 7 – 9), we study the *local identifiability* of l_1-minimization dictionary learning. We say that $\mathbf{D}_0$ is locally identifiable if it is a local minimum of the dictionary learning objective function. Instead of putting nonnegative constraints on both the dictionary and the coefficients, l_1-minimization dictionary learning derives the dictionary by minimizing the average l_1-norm of the linear coefficient vectors. Suppose we observe N data points $\mathbf{x}_i \in \mathbb{R}^K$ for $i = 1, ..., N$. The $\mathbf{x}_i$'s are $i.i.d.$ random linear combinations of the K columns from a complete (i.e., square and invertible) reference dictionary $\mathbf{D}_0 \in \mathbb{R}^{K \times K}$, where the random linear coefficients are generated from either the s-sparse Gaussian model or the Bernoulli-Gaussian model. For the population case in which we observe infinitely many data points, we established a sufficient and almost necessary condition for the reference dictionary $\mathbf{D}_0$ to be locally identifiable. Our condition characterizes the phase transition phenomenon of local identifiability and significantly improves the sufficient condition in Gribonval and Schnass (2010) [52]. For the finite sample case, we showed that similar local identifiability results hold with high probability if the number of samples $N = O(K \log K)$.

Since it is in general computationally expensive to check our sufficient and almost necessary condition, we also provided tight and easy-to-compute lower and upper bounds to approximate the quantities involved in the condition. With these bounds, we showed that for a complete μ-coherent reference dictionary, i.e., a dictionary with absolute pairwise column inner-product at most $\mu \in [0, 1)$, local identifiability holds even when the random linear coefficient vector has up to $O(\mu^{-2})$ nonzeros on average. Moreover, if the sparsity level is greater than $O(\mu^{-2})$, the reference dictionary is generally not locally identifiable. Our result is the first to show that $O(\mu^{-2})$ is both achievable and optimal for exact local recovery under the l_1-minimization criterion.

1.3 Organization of the thesis

This thesis is composed of two parts. Part I focuses on the analysis of spatial gene expression data. We will introduce our data and the preprocessing steps (Chapter 2), develop staNMF (Chapter 3), interpret the staNMF-derived dictionary (Chapter 4), perform gene categorization (Chapter 5) and build local gene networks (Chapter 6). Part II is devoted to theory – local identifiability of dictionary learning. We will give a detailed literature review and formulate the mathematical problem (Chapter 7), develop results in the population case (Chapter 8) and in the finite sample case (Chapter 9). Proofs can be found in the Appendix. We will conclude the thesis by describing future research directions for applications and theories (Chapter 10).

1.4 Datasets and software

Our data and code are available for download at http://insitu.fruitfly.org/downloads.

Part I

Analyzing spatial gene expression using stability-driven nonnegative matrix factorization

Chapter 2

Spatial gene expression patterns and data preprocessing

Identifying gene functions and gene-gene interactions is important for understanding human organogenesis and developmental diseases. As an animal egg develops from a single cell to an embryo with a full set of organ systems, combinations of genes are expressed in different spatial regions to establish body axis and trigger organ formation. Thus, for multicellular organisms, the study of spatial gene expression patterns is crucial for analyzing regulatory gene networks and gaining insight into organism development. Spatial expression profiling using *in situ* hybridization is one standard approach to systematically examine gene expression patterns. Large-scale embryonic mRNA expression pattern screens have been completed or are in progress for a number of model organisms [9, 15, 16, 17, 18, 19]. In this thesis, we used a set of *Drosophila* early embryonic spatial gene expression images, one of the largest datasets of its kind, from the ongoing Berkeley *Drosophila* Genome Project (BDGP) [9]. However, these images are not aligned and require registration prior to statistical analysis. Furthermore, gene expression data obtained through this approach contain certain imaging artifacts. In this chapter, we will address these issues by developing a data preprocessing pipeline.

2.1 Collecting *Drosophila* embryonic gene expression images

We generated a two-dimensional gene expression profile for *Drosophila melanogaster* during embryonic development. For each gene, RNA transcripts were detected by hybridization with an antisense DIG-labeled RNA probe and visualized using immunohistochemistry [9, 55, 21]. The blue stain in the embryo indicates where the gene is expressed (Figure 2.1). Images of the stained embryos were collected and manually classified into three orientations: lateral, dorsal and ventral, as well as six developmental stages: stages 1–3, 4–6, 7–8, 9–10, 11–12 and 13–16. Based on the collected images, a trained curator annotated each gene

with a controlled anatomical vocabulary [20]. In this thesis, we focused on lateral view embryos of developmental stages 4 – 6 (1 hour 20 min – 3 hours after egg laying at 25°C). Using the controlled vocabulary annotation, we removed images annotated solely with the terms "ubiquitous", "maternal" or "no staining". Compared to restricted zygotic expression patterns, these images had almost uniform expression intensities throughout the embryos and were therefore less important for our analysis. The resulting dataset contains 1640 images derived from 701 genes, 156 of which encode transcription factors (TF).

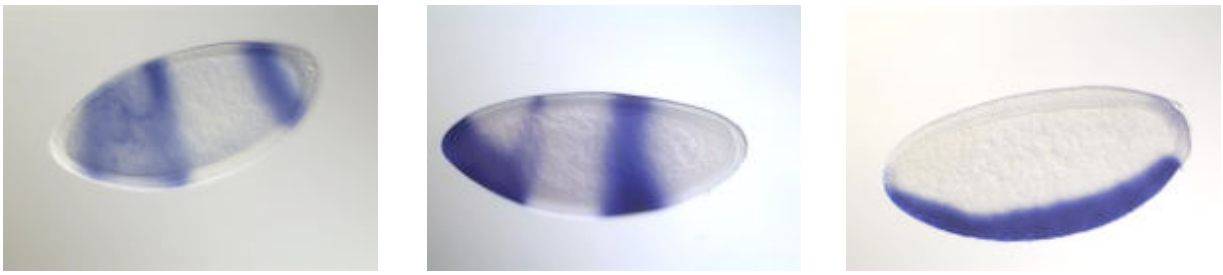

Figure 2.1: Stages 4–6 expression images for genes *hunchback (hb)*, *knirps (kni)* and *snail (sna)*. The blue stain indicates where the gene is expressed in the embryo.

2.2 Data preprocessing

Embryo registration

Individual *Drosophila* embryos vary in shape, size and orientation, and can also locate in different regions of the image (Figure 2.1). Thus, to meaningfully compare gene expression patterns, we registered each embryo onto a common template. For each image, we first detected the embryo outline using the segmentation algorithm in [27]. Next, we employed SPEX2 [28] to transform the near-elliptical embryo onto a standard ellipse template with long axis 64 pixels and short axis 32 pixels. In some cases, the embryo was flipped in the horizontal and/or vertical direction(s) such that the anterior part always faces left and the ventral part always faces down.

Expression pattern extraction

Our embryonic gene expression images were captured using differential interference contrast (DIC) microscopy. As a result, the shadows induced by DIC are frequently indistinguishable from expression patterns in grayscale [27]. We developed a least squares (LS) based method utilizing the color channels to differentiate spatial gene expression from background. Using Adobe® Photoshop®, we created a training set of 32 images by manually selecting the

regions of the embryos with gene expression, as detected by the blue dye. For each pixel inside the ellipse template, we averaged the three RGB values as proxy for gene expressing intensity and set the intensity outside the segmented region to zero. We then standardized the expression intensity g of each pixel using the formula $(255 - g)/255$. Here, 255 is the number of possible grayscale values. Under this standardization scheme, the maximum gene expression intensity is one and the minimum is zero.

For $i = 1, ..., 32$, we represented the i-th manually processed image as a vector $\mathbf{s}_i \in [0, 1]^{8192}$. The vector length 8192 was derived from 128×64 – the size of the rectangle image that contained the ellipse template. We predicted the gene expression intensity of each pixel as a linear combination of the color information of its neighbor pixel at different scales. Specifically, for each pixel and each color channel, we generated up to the fourth moment of the intensity at the pixel and within a disk centered at the pixel with radius 2, 4, and 8. Denote the feature matrix by $\mathbf{Z}_i \in \mathbb{R}^{8192 \times 49}$. Each row of $\mathbf{Z}_i$ corresponds to a pixel and each column a feature (1 column for the intercept, 4 moments $\times$ 4 radii $\times$ 3 channels = 48 features). Next, we estimated the linear coefficients using LS:

$$\hat{\mathbf{b}} = \arg \min_{\mathbf{b} \in \mathbb{R}^{49}} \sum_{i=1}^{32} \|\mathbf{s}_i - \mathbf{Z}_i \mathbf{b}\|_2^2.$$

The correlation between the predicted and the manually extracted gene expression is 0.9832. This number was quite high considering the fact that we had in our training data $8192 \times 32 = 262144$ pixels in total but only 48 features.

To extract the gene expression pattern for a new image, we first computed the feature vector $\mathbf{Z}_{new}$ and set $\mathbf{s}_{new} = \mathbf{Z}_{new}\hat{\mathbf{b}}$. Since the gene expression intensity should value between zero and one, we further truncated each entry of $\mathbf{s}_{new}$ to be in $[0, 1]$. We then used the resulting vector as the extracted gene expression pattern for further analysis. We evaluated our gene expression extraction procedure on a number of testing images. Our method performed well as indicated by the high correlation between the gene expression pattern extracted by the curator and the one predicted by the LS method (Figure 2.2).

Further downsampling and evaluation

The ellipse-registered embryo was further down-sampled to fit in an ellipse template with long axis 16 pixels and short axis 8 pixels. Such an ellipse template can be embedded in a rectangle image of 16×32 pixels for visualization. Inside the rectangle image, there are 405 pixels within the ellipse and $16 \times 32 - 405 = 107$ pixels outside. To validate our registration pipeline, we selected replicates of the same gene and genes with known adjacent expression patterns, superimposed them, and visually evaluated the matches to deem them satisfactory (Figure 2.3). See Figure 2.4 for a sample of preprocessed gene expression patterns.

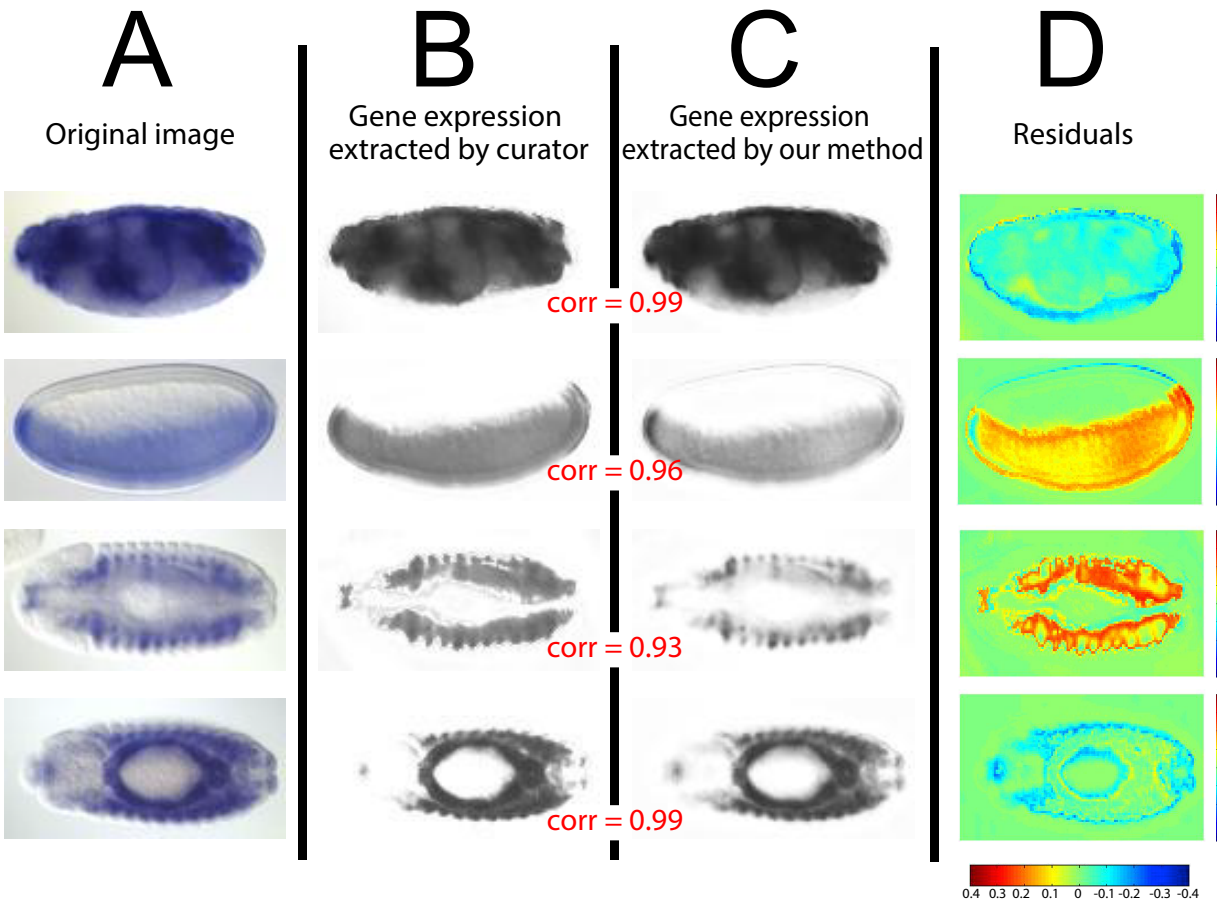

Figure 2.2: Extracting gene expression patterns from images obtained through differential interference contrast (DIC) microscopy. (**A**) The original image was standardized to an image of 64×128 pixels. (**B**) A curator used the selection tool in Adobe ® Photoshop® to extract regions of the *Drosophila* embryo deemed as having the blue dye. Since we averaged the three color channels to yield a proxy for gene expression intensity, this resulting image is displayed in gray scale, with white being the region of low expression, and black the region of high expression. (**C**) We extracted gene expression using a linear combination of the RGB features from the original image. For each example, the correlation indicated in red is the correlation between the gene expression extracted by curator and that extracted by our method. (**D**) The difference between the predicted pattern and the pattern extracted by the human curator is shown in the residual plot.

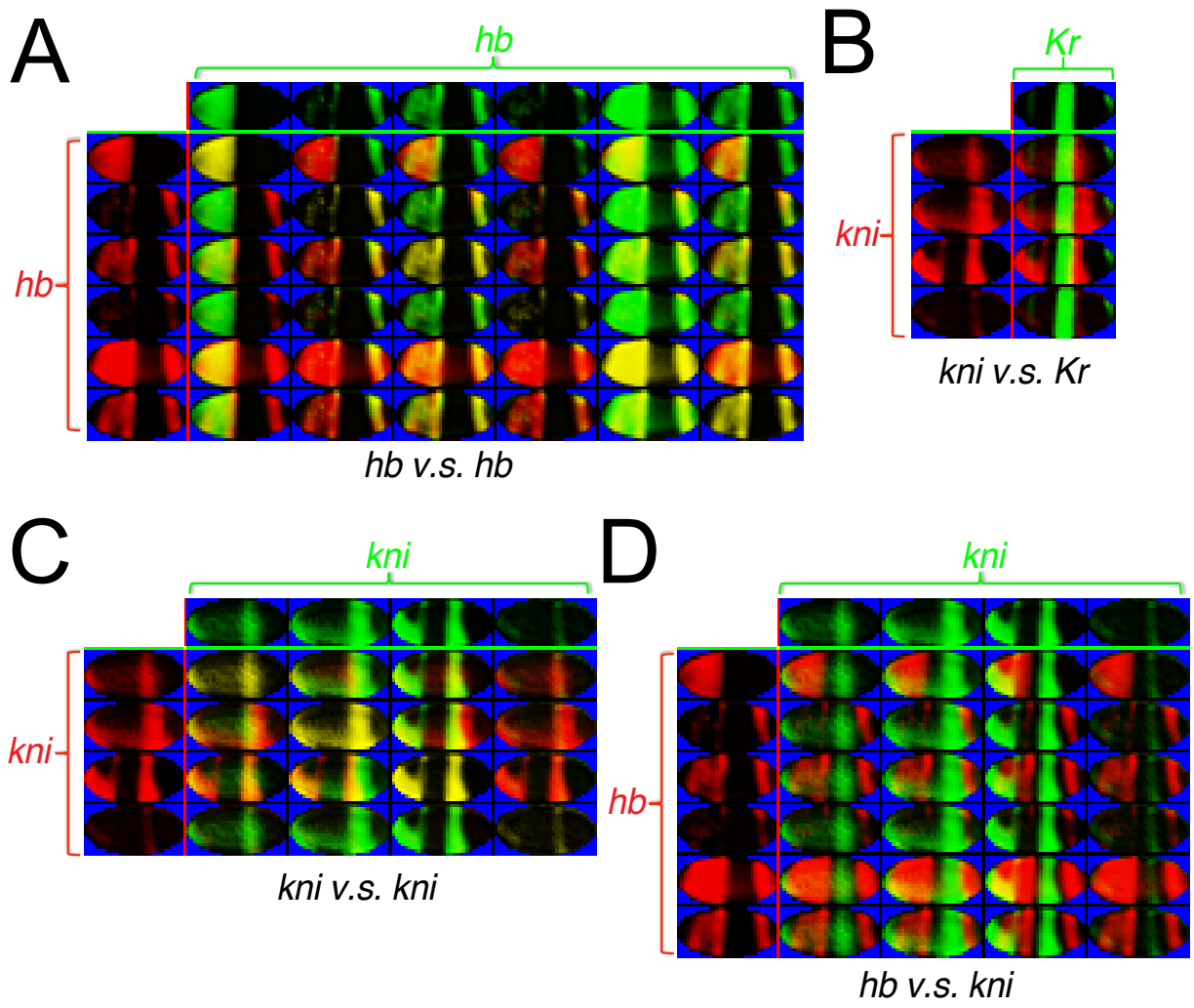

Figure 2.3: Visual evaluation of embryo registration. By overlaying two gene expression patterns with different colors, in this case, red and green, virtual double staining was performed between the replicates of the same gene (i.e., (**A**) *hb* v.s. *hb* and (**C**) *kni* v.s. *kni*), and between the replicates of genes one of which is known to be repressor of the other (i.e., (**B**) *kni* v.s. *Kr* and (**D**) *hb* v.s. *kni*). In both cases, the boundaries of the genes match, indicating that our registration approach performed reasonably well in transforming a *Drosophila* embryo into a common frame of reference.

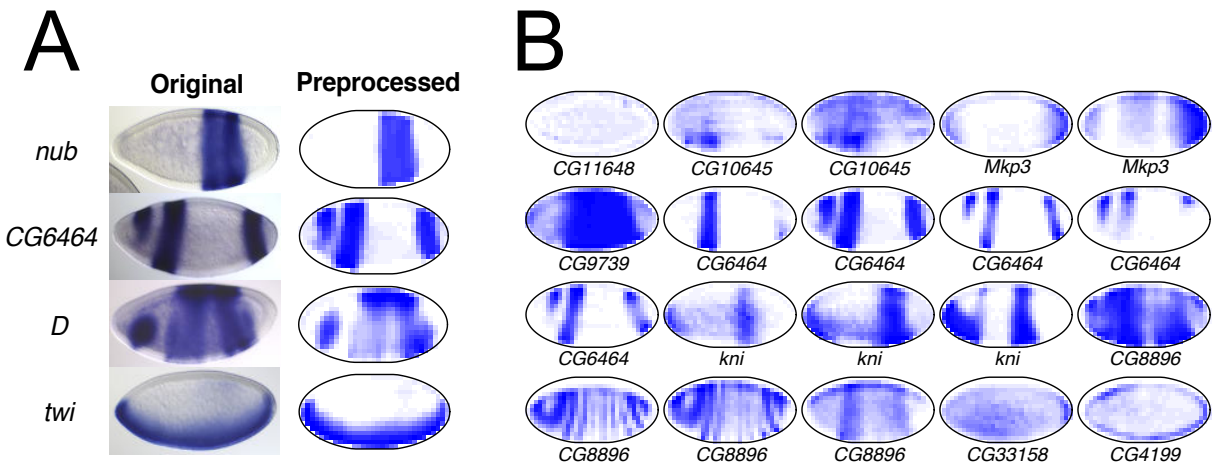

Figure 2.4: A sample of gene expression patterns in *Drosophila* embryos. (A) Expression patterns of four genes before and after the data preprocessing steps. (B) More examples of preprocessed gene expression patterns.

Chapter 3

Stability-driven nonnegative matrix factorization

In *Drosophila* development, cell fates are determined before any visible morphological features become apparent [35, 36]. They are preceded by the coordinated co-expression of cohorts of genes in defined spatial regions that divide the embryo into areas with unique regulatory profiles [27, 33]. Thus, we can think of each spatial gene expression as an additive and nonnegative linear combination of a set of regions of the embryo. In this chapter, we will describe a dictionary learning algorithm called nonnegative matrix factorization (NMF) [37], to identify these additive and positively valued regions. Following the stability principle [46], we designed a novel stability-based criterion to choose the number of dictionary atoms in NMF. Our method, called staNMF, performed very well in a number of simulation studies. When applied to the preprocessed *Drosophila* embryonic gene expression data, staNMF identified 21 biologically meaningful dictionary atoms, called principal patterns (PP). Based on the learned PP, we further utilized LASSO+NLS, a model selection and fitting procedure for nonnegative linear models, to provide compact representations for gene expression images.

3.1 NMF: formulation and algorithm

NMF is a popular unsupervised learning algorithm that can learn "parts-based" representation from the input data [37]. It has been used in many fields such as image processing and computer vision [39], text mining [40], audio signal separation [41] and bioinformatics [42]. See [43] for a recent review of NMF.

Denote by $\mathbb{R}_+$ the nonnegative real line. Let $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{d\times N}$ be the data matrix where each column represents a data vector. For a given positive integer K, NMF finds an entrywise nonnegative dictionary matrix $\mathbf{D} \in \mathbb{R}_+^{d\times K}$, under which each vector $\mathbf{x}_i$ has nonnegative representations: i.e., $\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i$ for a nonnegative vector $\boldsymbol{\alpha}_i \in \mathbb{R}_+^K$ (Figure 3.1). The nonnegativity constraints on both the dictionary and coefficients enforce the PP

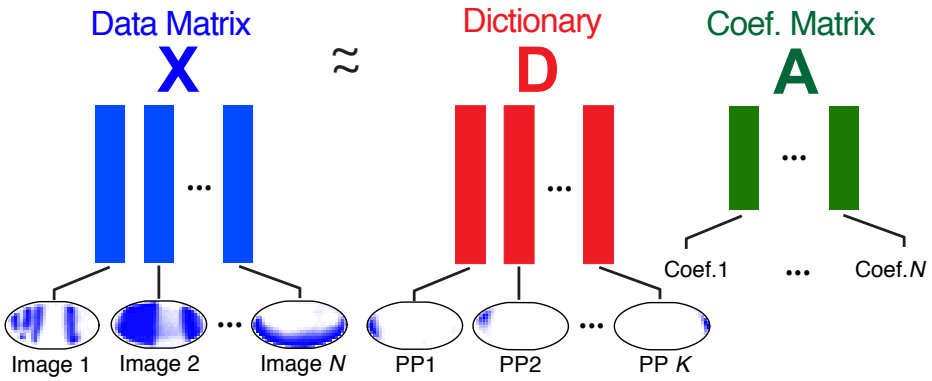

Figure 3.1: NMF on *Drosophila* embryonic gene expression image data. For a given number K, NMF factorizes the nonnegative data matrix $\mathbf{X}$, the columns of which are gene expression images, into the product of two nonnegative matrices: dictionary $\mathbf{D}$, which contains the K PP, and coefficient matrix $\mathbf{A}$, which contains the nonnegative coefficients of the images.

to have nonnegative contributions to the observations, resulting in a parts-based representation for image applications. Mathematically, NMF aims at solving the following nonconvex optimization problem:

$$\min_{\mathbf{D},\mathbf{A}=[\boldsymbol{\alpha}_1,...,\boldsymbol{\alpha}_N]} \|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2 = \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2,$$
$$\text{subject to } \mathbf{D} \geq 0, \|\mathbf{D}[,k]\|_2 \leq 1 \text{ for } k = 1,..,K,$$
$$\text{and } \boldsymbol{\alpha}_i \geq 0, \text{ for } i = 1,...,N.$$

Here, $\mathbf{D}[,k]$ is the k-th column of the dictionary $\mathbf{D}$. Note that the above formulation of NMF does not require the data matrix $\mathbf{X}$ to be nonnegative in every entry. For some numerical examples, $\mathbf{X}$ is the product of two nonnegative matrices contaminated with noise and hence can be negative in some entries (see Simulation Experiment 1 in the below section).

For our *Drosophila* gene expression data, $\mathbf{x}_i$ is a vector of length 405 that corresponds to the i-th preprocessed spatial gene expression pattern. To account for possible replicates of the same gene, we used a weighted version of NMF with the following modified objective function:

$$\sum_{i=1}^{N} \mathbf{w}[i]\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2,$$

where the weight for the i-th image $\mathbf{w}[i]$ is the reciprocal of the number of replicates of the gene that corresponds to the i-th image (from now on, we will denote the j-th entry of a vector $\mathbf{v} \in \mathbb{R}^m$ as $\mathbf{v}[j]$). Note that the above objective function can be rewritten as $\mathbf{w}[i]\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 = \|\sqrt{\mathbf{w}[i]}\mathbf{x}_i - \mathbf{D}(\sqrt{\mathbf{w}[i]}\boldsymbol{\alpha}_i)\|_2^2$. Therefore we can simply set $\mathbf{x}'_i = \sqrt{\mathbf{w}[i]}\mathbf{x}_i$ and use any algorithm that solves original NMF formulation, with $\mathbf{X}' = [\mathbf{x}'_1, ..., \mathbf{x}'_N]$ as the

new data matrix. Denote by $(\hat{\mathbf{D}}, \hat{\mathbf{A}}')$ the output of the NMF algorithm. The nonnegative coefficient matrix $\hat{\mathbf{A}}$ can be retrieved by scaling the i-th column of the matrix $\hat{\mathbf{A}}'$ by the factor $\mathbf{w}[i]^{-1/2}$.

To compute NMF, we used the `SPAMS` package with the MATLAB interface [47]. `SPAMS` implemented a number of online algorithms for dictionary learning and matrix factorization. The package is fast and scales to large numbers of data points. The NMF algorithm requires an initial input dictionary. We constructed this input by randomly sampling K columns from the data matrix $\mathbf{X}$. To compute the dictionary, `SPAMS` performed alternating minimization: given the current iteration of the dictionary $\mathbf{D}$, update the nonnegative coefficients $\boldsymbol{\alpha}_i$'s using nonnegative least squares (NLS); and given the nonnegative coefficients, update the dictionary $\mathbf{D}$ by solving another series of NLS. We ran the algorithm until convergence. Clearly, the output dictionary $\hat{\mathbf{D}}$ depends on the initial input. This property can be further exploited to choose the number of dictionary atoms, as explained in the below section.

3.2 staNMF: stability-driven NMF model selection

In this section, we will address the issue of choosing the number K of dictionary atoms in NMF. As mentioned, since `SPAMS` solves NMF by an alternating minimization algorithm, the output dictionary depends on the initial value. We reasoned that a useful definition of an optimal NMF-generated dictionary would be reproducibly independent of the initialization values. We proposed staNMF, a procedure that combined multiple runs of NMF with a new Amari-type criterion to measure the instability of output dictionaries, to perform model selection in NMF.

For each K, we ran the NMF algorithm B times. Typically, $B = 100$ for the *Drosophila* gene expression data and other simulated examples presented in the thesis. For each NMF run, the columns of the initial dictionary were randomly sampled (without replacement) from the columns of $\mathbf{X}$. The B NMF runs generated output dictionaries $\hat{\mathbf{D}}_b$ for $b = 1, ..., B$.

Next, we will introduce a dissimilarity measure for two dictionaries of the same matrix dimension. Let $\mathbf{C} \in \mathbb{R}^{K \times K}$ be the cross correlation matrix between the atoms of two dictionaries $\mathbf{D}_1$ and $\mathbf{D}_2$ with the same number K of atoms. For a matrix $\mathbf{H} \in \mathbb{R}^{m \times n}$, denote by $\mathbf{H}[j,k]$ its (j,k)-th entry. Since the columns of a dictionary are permutation invariant, to measure dissimilarity between $\mathbf{D}_1$ and $\mathbf{D}_2$, we designed the following Amari-type quantity:

$$\begin{aligned} diss(\mathbf{D}_1, \mathbf{D}_2) &= \frac{1}{2}\left(\frac{1}{K}\sum_{j=1}^{K}\left(1 - \max_{1\le k\le K}\mathbf{C}[k,j]\right) + \frac{1}{K}\sum_{k=1}^{K}\left(1 - \max_{1\le j\le K}\mathbf{C}[k,j]\right)\right) \\ &= \frac{1}{2K}\left(2K - \sum_{j=1}^{K}\max_{1\le k\le K}\mathbf{C}[k,j] - \sum_{k=1}^{K}\max_{1\le j\le K}\mathbf{C}[k,j]\right). \end{aligned}$$

Note that when $\mathbf{D}_2$ can be transformed into $\mathbf{D}_1$ by column permutation, $diss(\mathbf{D}_1, \mathbf{D}_2) = 0$. Such a definition was inspired by Amari et al. [56], who used a comparable quantity to

measure the performance of their blind signal separation algorithm. The discrepancy of all B dictionaries for K was measured by the average Amari-type error of all $B(B-1)/2$ pairs of dictionaries:

$$\Upsilon(K) = \frac{2}{B(B-1)} \sum_{1 \le b < b' \le B} diss(\hat{\mathbf{D}}_b, \hat{\mathbf{D}}_{b'}).$$

We selected K that achieved a small $\Upsilon(K)$, i.e., a small discrepancy or instability. Once the parameter K was determined, we selected the learned dictionary with the minimum NMF square loss among all B dictionaries.

Brunet et al.'s stability-based criterion

The idea of using stability for NMF model selection was first introduced by Brunet et al. [42]. However, their stability metric was substantially different from ours. In their paper, NMF was used for cluster analysis. They proposed to choose K such that their NMF cluster assignment is most stable. Given a dictionary $\mathbf{D}$ with K columns, they assigned the data vector $\mathbf{x}_i$ to the k-th cluster, if the nonnegative coefficient for the k-th dictionary atom has the highest value among all K coefficients. If more than one dictionary atom share the same coefficient value, the data point is assigned to any of the corresponding clusters with equal probability. For the clustering defined by NMF, they constructed the connectivity matrix S, whose (i, j)-th entry is set to one if the i-th and the j-th data points belong to the same cluster, and zero otherwise. Based on the B NMF runs, they computed the consensus matrix, $\bar{S}$, which was defined as the average of all connectivity matrices. They then used the cophenetic correlation coefficient based on $\bar{S}$ to measure the clustering stability of NMF. In their paper, the cophenetic correlation coefficient was defined as the Pearson correlation coefficient of (1) the distance between the i-th and j-th data points as measured by $1-\bar{S}[i, j]$ and (2) the distance between the i-th and j-th data points induced by the average linkage hierarchical clustering using $\bar{S}$ as the similarity matrix, for all $1 \le i < j \le N$ (recall that N is the number of data points). The closer the cophenetic correlation coefficient to 1, the more stable the clustering assignment. To compare with our method, we used the equivalent one minus the cophenetic correlation coefficient, which is now a measure for clustering instability, and strived for a minimum value.

Synthetic and real data applications

We tested our staNMF as well as Brunet et al.'s method on a number of synthetic data with a known ground truth dictionary. While both methods identified the same K for some examples (Simulation Experiment 1 and 3), it is not surprising that Brunet et al.'s method failed on the others (e.g., Simulation Experiment 2), as their method was originally designed for the purpose of cluster analysis. Our staNMF performed consistently well. When applied to our *Drosophila* spatial gene expression data, both stability-based methods agreed on $K = 21$. Below, we will describe our simulation experiments and the real data application.

Simulation Experiment 1

In this experiment, we investigated how the two stability-based methods behave for dictionaries with different coherence and linear coefficients with various sparsity. It has been shown that increased dictionary coherence, or collinearity between dictionary atoms, might lead to ill-posedness of a number of dictionary learning formulations, see e.g., [52, 53, 54] and Part II of this thesis. Empirically, we also found it difficult for NMF to recover the dictionary if the atoms were highly collinear. Therefore, as the coherence of the dictionary increases, we suspected that it is increasingly challenging for the two stability-based methods to identify the correct K.

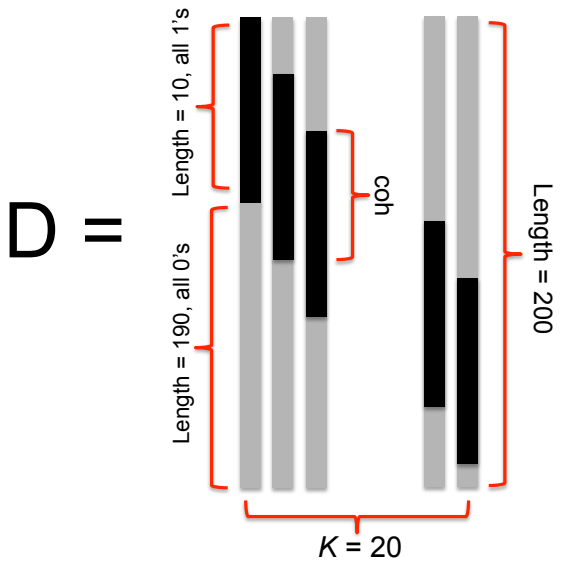

Figure 3.2: Construction of the dictionary in Simulation Experiment 1. In the above illustration, each vertical bar represents a dictionary atom (column). The black region of a bar indicates the entries that are ones and gray region the entries that are zeros. The parameter $coh \in \{0, 1, ..., 10\}$ is the number of common entries that are ones between two consecutive dictionary atoms. It measures also the coherence of the dictionary. From the i-th bar to the $(i+1)$-th bar, the black region is shifted down by the constant amount $10 - coh$.

We generated our data as follows. Let $coh \in \{0, 1, ..., 10\}$. We constructed the ground truth dictionary $\mathbf{D}_0 \in \mathbb{R}^{200\times 20}$ as:

$$\mathbf{D}_0[j,k] = \begin{cases} 1, & \text{for } 1 + (k-1)(10-coh) \leq j \leq 10 + (k-1)(10-coh), \\ 0, & \text{otherwise.} \end{cases}$$

See Figure 3.2 for an illustration of the above dictionary construction. Under this construction, each dictionary atom has exactly 10 entries equal to one and the remaining 190

entries equal to zero. Furthermore, two consecutive dictionary atoms share *coh* entries that are equal to one in common. Thus, the parameter *coh* controls the coherence of the dictionary, which is defined as the maximum absolute inner product between dictionary atoms.

Next, we generated the entries of the coefficient matrix $\mathbf{A}_0 \in \mathbb{R}^{20\times 1000}$ as independent and identical Bernoulli random variable with success probability $0 < p \leq 1$. Set the data matrix $\mathbf{X} = \mathbf{D}_0\mathbf{A}_0 + \mathbf{E}$, where $\mathbf{E} \in \mathbb{R}^{200\times 1000}$ was a noise matrix with entries drawn independently and identically from a Gaussian distribution with mean zero and standard deviation 0.1.

For each combination of (p, coh), we ran NMF $B = 100$ times for $10 \leq K \leq 30$ and then applied both stability criteria. The results shown in the Figure 3.3 indicated that when the dictionary coherence was low, our measure for dictionary instability, $\Upsilon(K)$, had a clear minimum at $K = 20$ which was the true number of dictionary atoms. However, as the dictionary coherence increased, for example, $coh = 6$, $\Upsilon(K)$ as a function of K changed shape and multiple local minima with similar stability emerged. This observation supported our previous conjecture that a higher dictionary coherence made it more difficult for the staNMF to identify the correct K. It is unclear how the sparsity parameter p affected our stability criterion.

Brunet et al.'s method behaved similar on the same data (Figure 3.4). However, we found that the clustering instability measure was versatile across the range of K and had too many abrupt local minima. On the other hand, our measure of dictionary instability $\Upsilon(K)$ was much more continuous and predictable. For example, for $p = 1$ and $coh = 2$, Brunet et al.'s stability curve had two almost identical local minima: one at $K = 10$ and the other at $K = 20$. In this case, their method was not robust: slight contamination of the data might mislead their method to choose $K = 10$ as the best number of dictionary atoms. For the same example, staNMF gave a very clear minimum at $K = 20$.

Simulation Experiment 2: the Swimmer data

In this example, we evaluated staNMF with a dataset that has been widely used in the NMF literature: the Swimmer data [57, 44, 45]. The dataset contained 256 images each of 32×32 pixels depicting all possible gestures of an artificial swimmer (Figure 3.5*A* and *B*). For each image, each limb of the swimmer was chosen from one of four gestures for that limb. The true dictionary therefore consisted of $4 \times 4 = 16$ atoms and so the number of all possible combinations of the swimmer gestures was $4^4 = 256$.

For this data, our method recovered the correct $K = 16$ (Figure 3.5*C*). However, Brunet et al.'s method chose $K = 14$ (Figure 3.5*D*). To elucidate the reason, we noted that each swimmer image had equal contribution from four dictionary atoms. Thus under the ground truth dictionary, each image should be assigned to the corresponding four clusters simultaneously. However, Brunet et al.'s approach forced the image to belong to only one cluster. As a result, it would select any one of the four clusters with equal probability. The randomness of an image falling into one of the four clusters resulted into clustering instability at $K = 16$. In contrast, staNMF did not assume any clustering structure and so it also identified the correct K for this dataset.

Real Data: *Drosophila* gene expression patterns

For each of the 1640 *Drosophila* gene expression images, we converted the pixel intensities of the preprocessed expression pattern into a linear vector of length 405 and decomposed the vector with NMF. We applied both stability-based criteria to the data for $15 \leq K \leq 30$. The dictionary learned with $K < 15$ resulted in PP that were in general too broad, as compared to the pre-organ partitions in the *Drosophila* fate map. These PP also led to poor reconstruction quality when using them to represent the gene expression patterns. Dictionary learned with $K > 30$ resulted in PP that were too unstable. For $15 \leq K \leq 30$, both staNMF and Brunet et al.'s method identified $K = 21$ as the optimal number of PP (Figure 3.6A).

We noticed that our stability criterion, $\Upsilon(K)$ had similar values for $K = 21$ and $K = 22$. When comparing the $K = 22$ dictionary with the $K = 21$ dictionary, we found that three PP from the $K = 22$ dictionary, PP4, PP5 and PP6, were different from the corresponding PP4 and PP5 of the $K = 21$ dictionary (Figure 3.7). In particular, PP5 in the $K = 21$ dictionary was split into PP5 and PP6 in the $K = 22$ dictionary. The remaining 19 PP were essentially unchanged. Thus the PP learned using the two different K were very similar. For simplicity we chose $K = 21$.

Simulation Experiment 3: the denoised *Drosophila* data

In Simulation Experiment 1, we demonstrated that dictionary coherence might affect the two stability-based model selection criteria. As a sanity check for our real data application, we generated a dataset using the 21 PP learned from the *Drosophila* data and investigated whether staNMF can recover the correct number of PP from this artificial data. Specifically, denote by $\hat{\mathbf{D}} \in \mathbb{R}_+^{405\times 21}$ the learned dictionary which contains the 21 PP and $\hat{\mathbf{A}} \in \mathbb{R}_+^{21\times 1640}$ the corresponding nonnegative coefficient matrix. We generated the data matrix $\hat{\mathbf{X}} \in \mathbb{R}_+^{405\times 1640}$ as the "denoised" version of the original data matrix: $\hat{\mathbf{X}} = \hat{\mathbf{D}}\hat{\mathbf{A}}$. For this dataset, both staNMF and Brunet et al.'s method selected $K = 21$ as the optimal number of PP (Figure 3.6B).

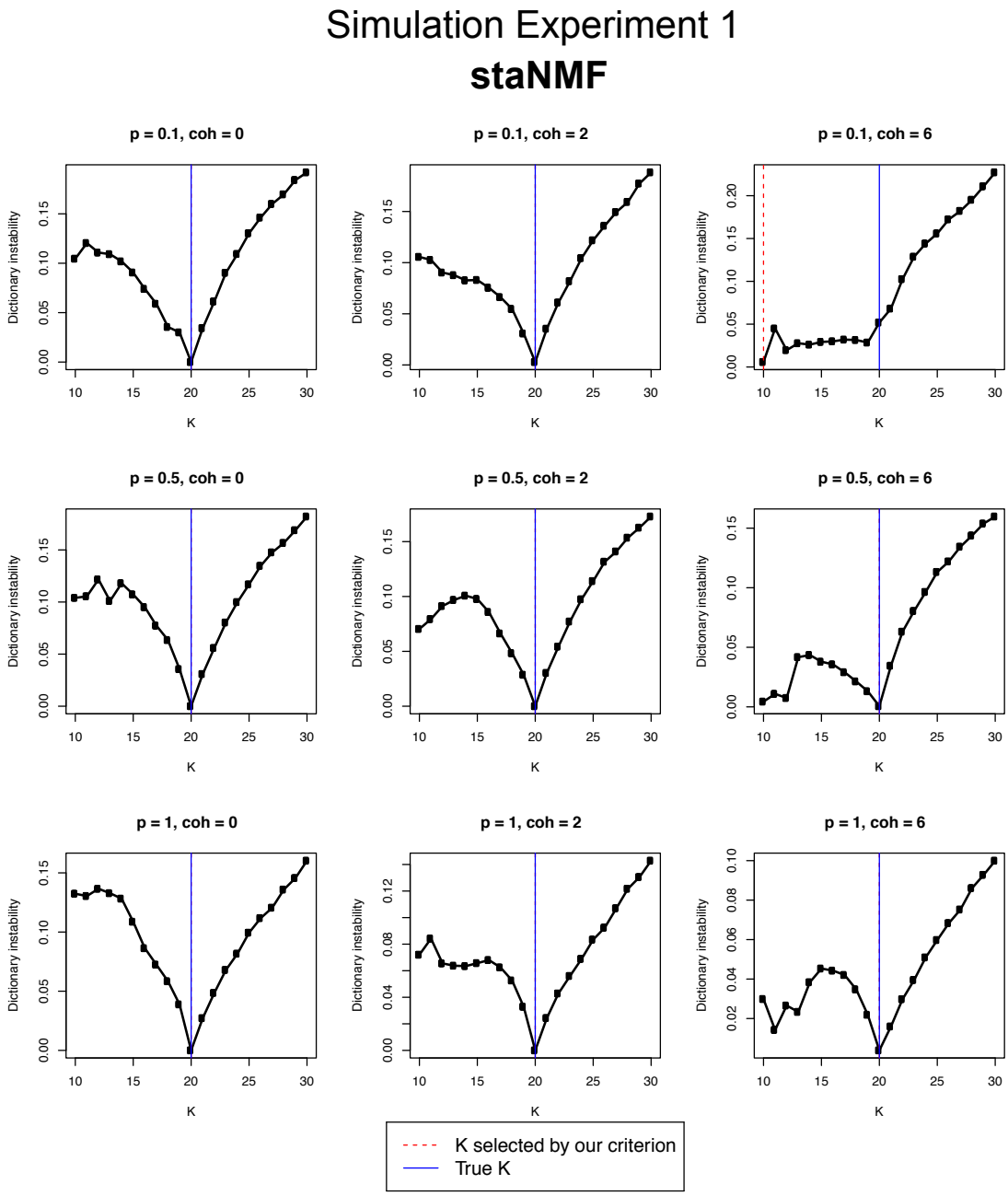

Figure 3.3: NMF model selection using staNMF: Simulation Experiment 1. For each $(p, coh) \in \{0.1, 0.5, 1\} \times \{0, 2, 6\}$, we generated the data matrix $\mathbf{X}$ from the model described in Simulation Experiment 1. For each parameter configuration, we ran NMF $B = 100$ times for every $10 \leq K \leq 30$ and then applied our stability criterion. For each plot, the vertical axis represents the dictionary instability as measured by $\Upsilon(K)$ defined in the text. The lower the value, the more stable the dictionaries with respect to random initial values.

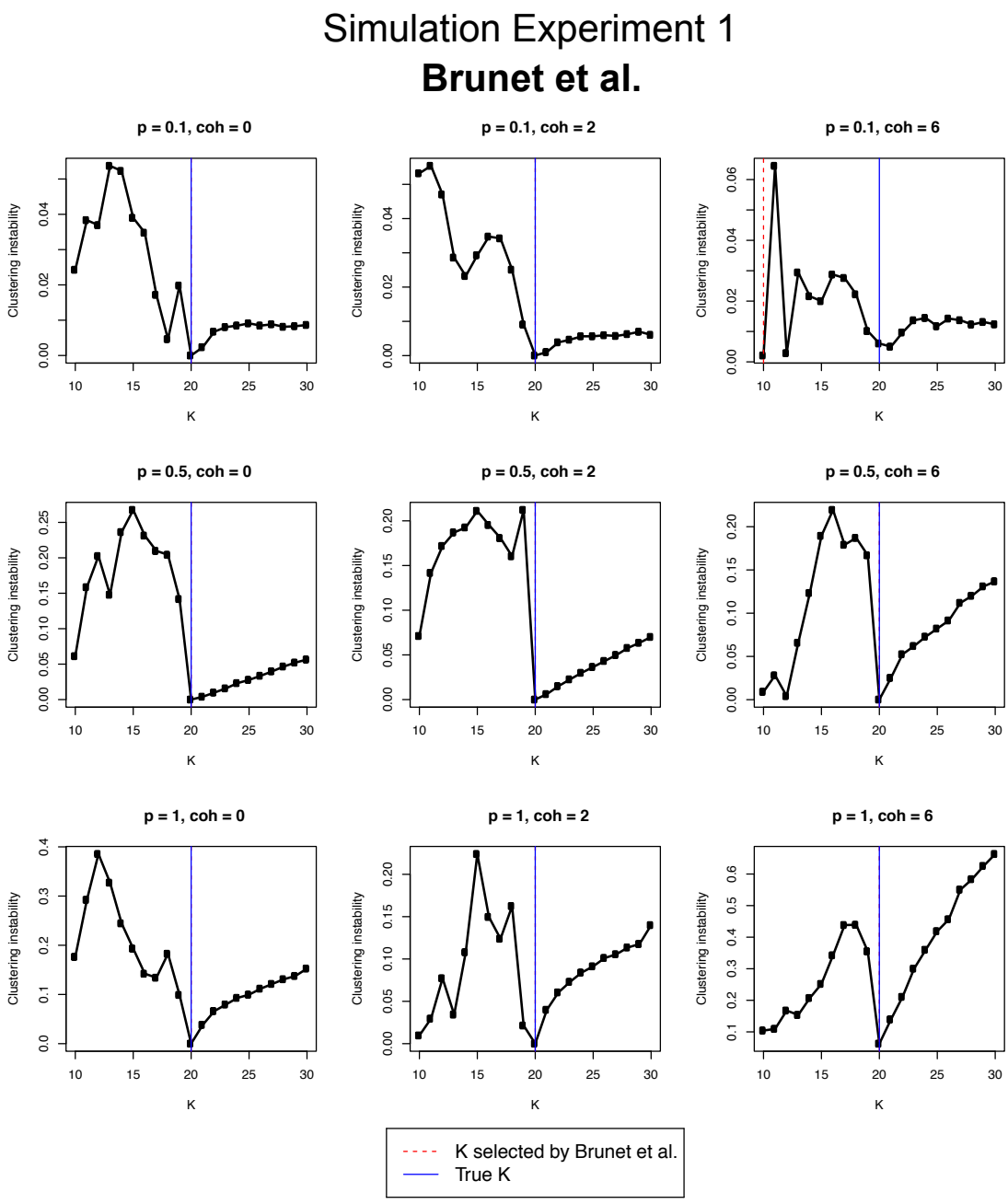

Figure 3.4: NMF model selection using Brunet et al.'s clustering instability criterion [42]: Simulation Experiment 1. The synthetic data are exactly the same as those in Figure 3.3. For each plot, the vertical axis represents the clustering instability as measured by one minus the cophenetic correlation coefficient of the NMF cluster consensus matrix (see text). The lower the value, the more stable the cluster assignment with respect to random initial values.

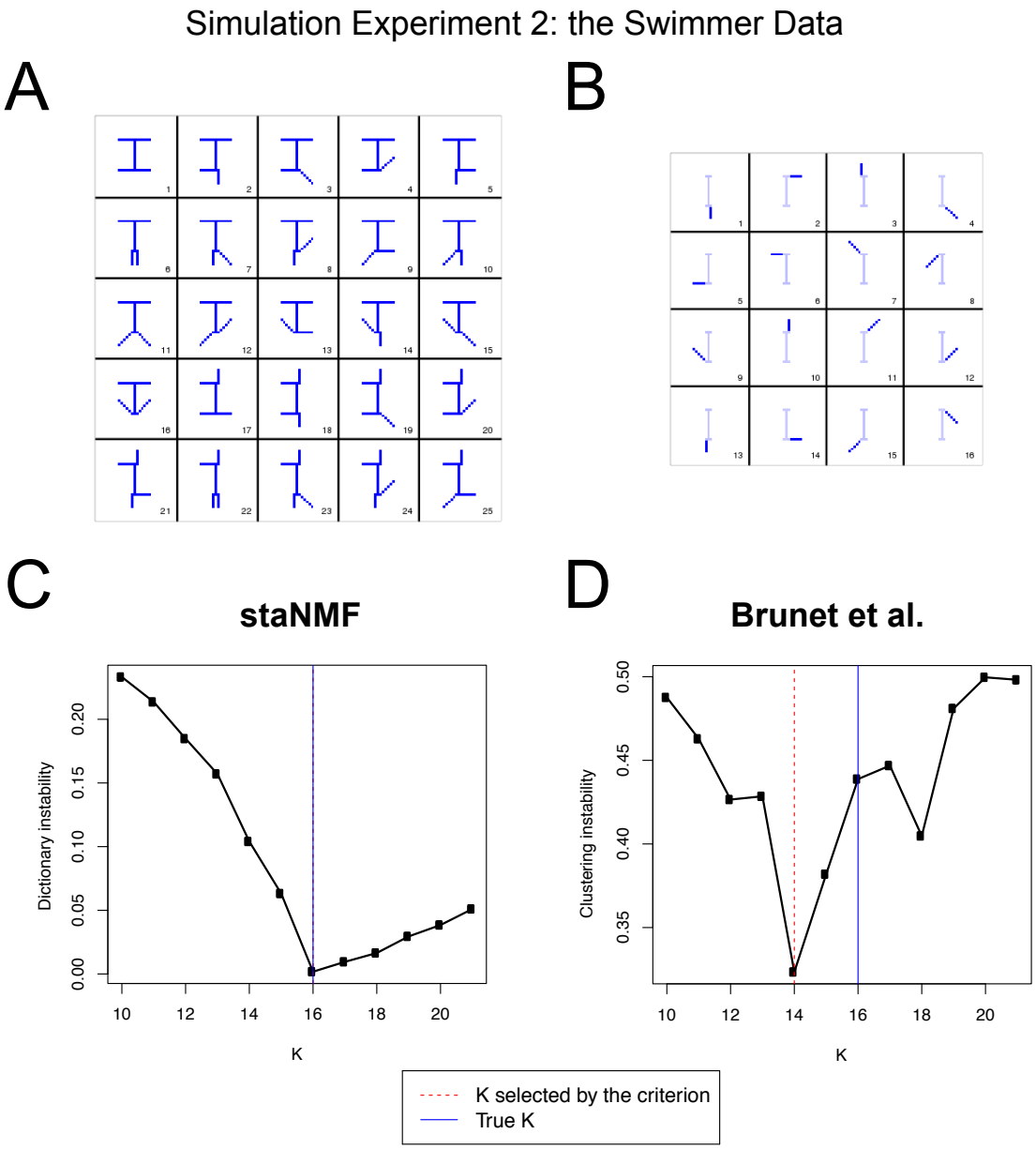

Figure 3.5: NMF model selection: Simulation Experiment 2 – the Swimmer dataset [57]. (**A**) A sample of 25 images containing the artificial swimmers. (**B**) The 16 dictionary atoms recovered by NMF. The dark blue region of each basis image corresponds to a limb of the artificial swimmer, whereas the light blue region indicates the torso of the swimmer. (**C**) staNMF identified correctly $K = 16$. (**D**) Brunet et al.'s method selected $K = 14$.

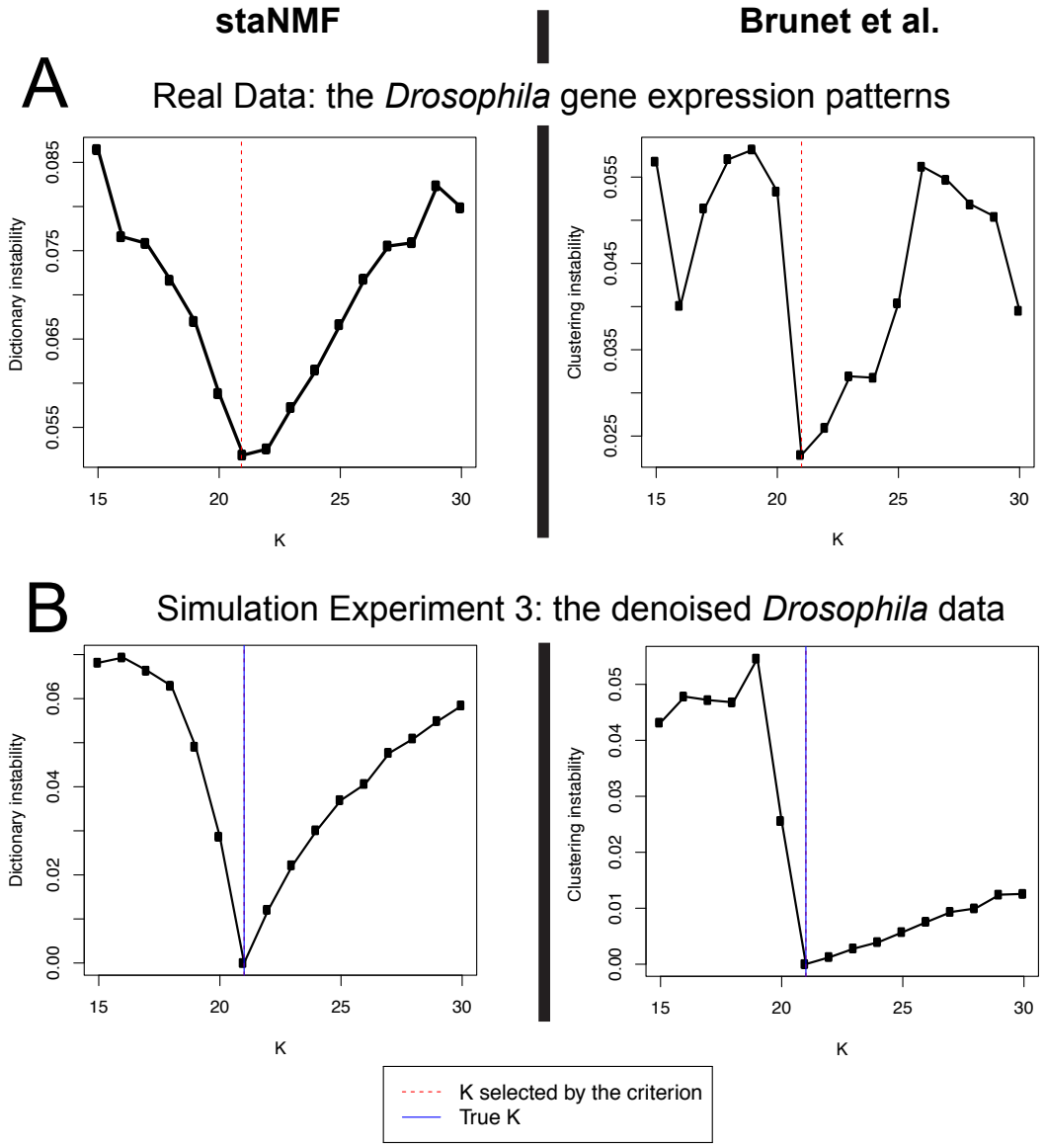

Figure 3.6: staNMF and Brunet et al.'s stability criterion on (**A**) the *Drosophila* spatial gene expression data and (**B**) the corresponding denoised data. The two methods agreed on $K = 21$ in both examples. Note that for the real data we do not know the true number of PP and so only the red dash lines were drawn for the two plots in (A). The denoised data was constructed as $\hat{\mathbf{X}} = \hat{\mathbf{D}}\hat{\mathbf{A}}$, where $(\hat{\mathbf{D}}, \hat{\mathbf{A}})$ is the output dictionary and the nonnegative coefficient matrix from the NMF.

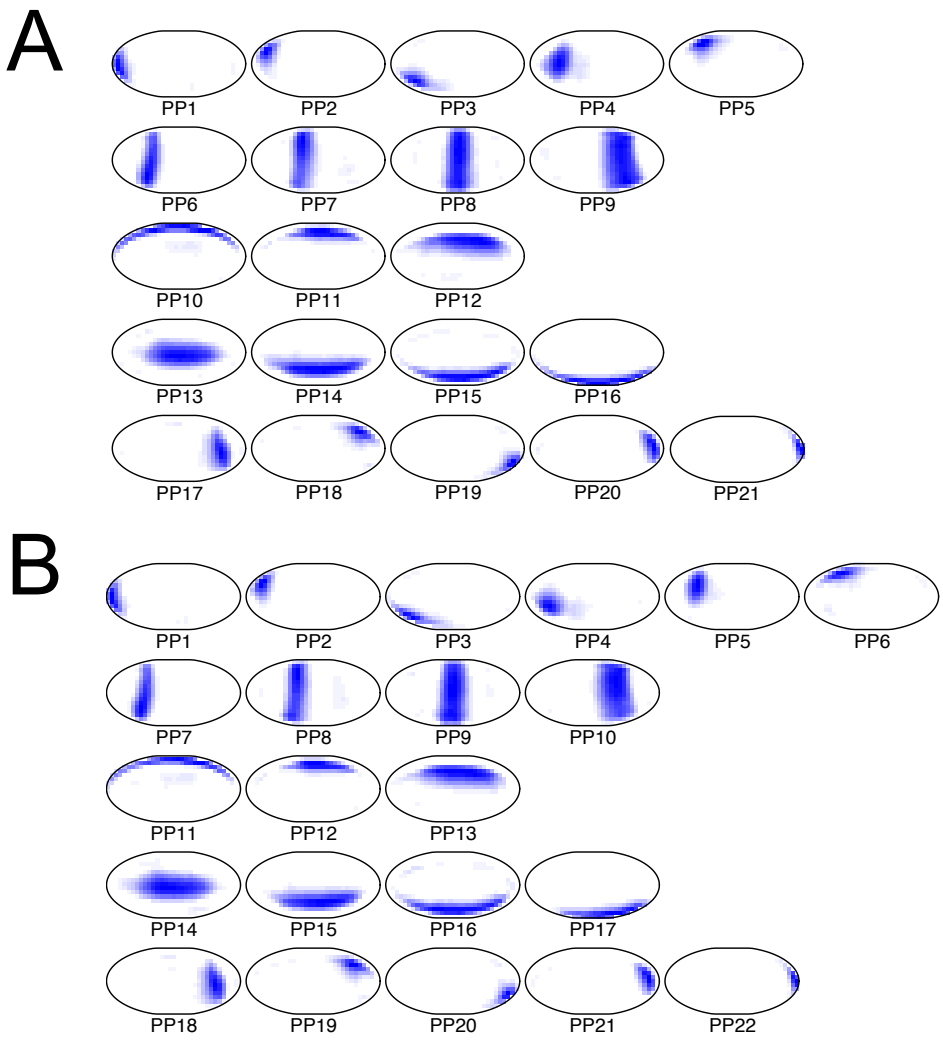

Figure 3.7: NMF dictionaries learned with number of dictionary atoms (**A**) $K = 21$ and (**B**) $K = 22$. Every PP was normalized to have maximum intensity equal to one.

3.3 Representing spatial expression patterns by the learned PP

We evaluated the ability of PP to provide a compact representation for spatial gene expression patterns. A sparse decomposition of complex expression patterns into additive smaller components offers a simple and intuitive computational representation of spatial gene expression. Using nonnegative least squares (NLS), the NMF algorithm gave a nonnegative linear representation matrix of the data under the learned dictionary. With the nonnegativity as an implicit sparsity penalty, NLS can be treated as a method to perform model selection [58]. However, empirically we found that NLS selected more covariates than necessary. To address this issue, we added a LASSO step before the NLS and showed that this strategy outperformed the plain NLS approach.

NLS tends to over-select covariates

We generated 1000 data vectors using the model discussed in Simulation Experiment 1 with dictionary coherence $coh = 2$ and random linear coefficient sparsity $p = 0.2$. For each of the data vector generated, we applied NLS to estimate the nonnegative linear coefficients with the ground truth dictionary as the covariate matrix. The resulting NLS coefficients contained many more nonzeros than the true nonnegative linear coefficients used to generated the data (Figure 3.8A and B). The average support difference between the estimated coefficient and the true coefficient was 7.29, out of a maximum of 40.

The LASSO+NLS procedure for model selection and fitting

To address the above issue, we employed the following LASSO+NLS procedure. Let $\mathbf{x} \in \mathbb{R}^d$ be a data vector and $\mathbf{D} \in \mathbb{R}^{d \times K}$ the dictionary or covariate matrix. We first used the LASSO, or least absolute shrinkage and selection operator [48], with the nonnegative constraints on the linear coefficients:

$$(\hat{\mu}, \hat{\beta}(\lambda)) = \arg \min_{\mu \in \mathbb{R}_+, \beta \in \mathbb{R}_+^K} \|\mathbf{x} - \mathbf{D}\beta - \mu\|_2^2 + \lambda \|\beta\|_1.$$

With a 10-fold cross-validation, the LASSO regularization parameter λ was chosen to be the largest among all parameters whose cross-validation error was within one standard error of the minimum cross-validation error. Denote by $\hat{\beta}_{lasso}$ the nonnegative linear coefficient at the selected λ.

Due to the l_1-penalty term, the LASSO estimator is biased towards zero for finite samples. In order to reduce the bias, we fitted NLS on the dictionary atoms selected by the LASSO [59]. Let $S = \{k : \hat{\beta}_{lasso}[k] \neq 0\} \subset \{1, ..., K\}$ be the support of the LASSO coefficient vector and $\mathbf{D}[, S]$ be the submatrix of $\mathbf{D}$ with columns indexed by S. We solved the following NLS problem:

$$(\hat{\nu}, \hat{\gamma}) = \arg \min_{\nu \in \mathbb{R}_+, \gamma \in \mathbb{R}_+^{|S|}} \|\mathbf{x} - \mathbf{D}[, S]\gamma - \nu\|_2^2,$$

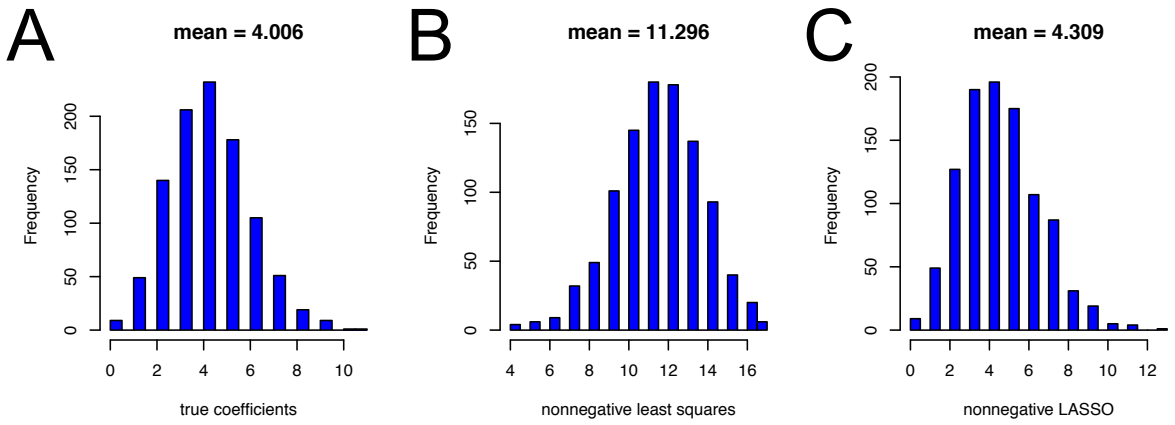

Figure 3.8: Effectiveness of the LASSO+NLS model selection and fitting procedure. We generated 1000 data vectors according to the model described in Simulation Experiment 1, with dictionary parameter $coh = 2$ and random linear coefficient sparsity $p = 0.2$. For each data vector, using the ground truth dictionary as the covariate matrix, both nonnegative least squares (NLS) and LASSO+NLS were applied to estimate the linear coefficients. Shown are the histograms of the number of nonzeros in the linear coefficients of (**A**) the true model, (**B**) the NLS estimates and (**C**) the LASSO+NLS estimates. If model selection is performed properly, the resulting distribution of number of nonzeros should match with that of the true coefficients. Here, the distribution of the number of nonzeros for NLS shifted significantly to the right (B), indicating that NLS tended to over-select covariates. The LASSO+NLS fitting procedure, on the other hand, produced number of nonzeros distribution almost identical to histogram for the true coefficients (C).

where $|S|$ is the size of the set S. The *sparse PP (sPP) representation* or *sPP coefficient* for the data vector $\mathbf{x}$, denote by $\eta \in \mathbb{R}_+^K$, is a vector whose entries indexed by S, $\eta[S] = \hat{\gamma}$ and entries indexed by the complement of S, $\eta[S^c] = 0$.

We applied the LASSO+NLS procedure to our previous simulation example. The distribution of number of nonzero estimated coefficients per observation now matched that of the true coefficients (Figure 3.8*C*). The average support difference between the two reduced significantly to 0.3.

We used the R package `glmnet` [60] for the computation.

Application to our data

We applied this PP selection and fitting procedure to our data (Figure 3.10). The average number of PP chosen by this procedure is 10.4, and the average correlation between the original expression pattern and the reconstructed pattern is 0.854 (Figure 3.9*A* and *B*). Considering the small number of the selected PP, the correlation measure indicates that

our model selection and fitting procedure achieved a reasonably good reconstruction quality. As expected, the correlation increases as the number of PP increases (Figure 3.9*C*). We investigated cases with poor performance and found such gene expression patterns are either faint or with poorly defined boundaries. In addition, non-sparse representations almost always correspond to ubiquitously expressed genes (Figure 3.9*D*). As illustrated by the residual images, errors are most likely to occur at expression pattern boundaries (Figure 3.10).

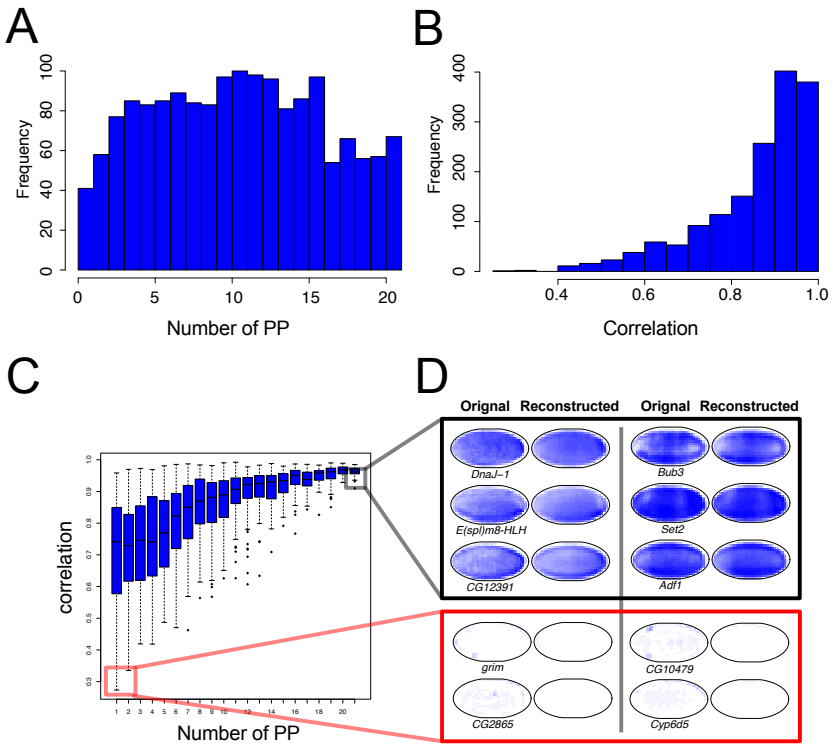

Figure 3.9: Spatial gene expression reconstruction quality of the sparse PP (sPP) representation. (**A**) Histogram of the number of selected PP per expression pattern. (**B**) Histogram of the correlation between a gene expression pattern and the reconstructed pattern (linear combination of the 21 learned PP using the sPP representation as coefficients). (**C**) The relationship between the number of selected PP and the correlation. (**D**) A sample of expression patterns represented with more than 20 PP (black box) and a sample of expression patterns with poor reconstruction quality, i.e., correlation less than 0.35 (red box).

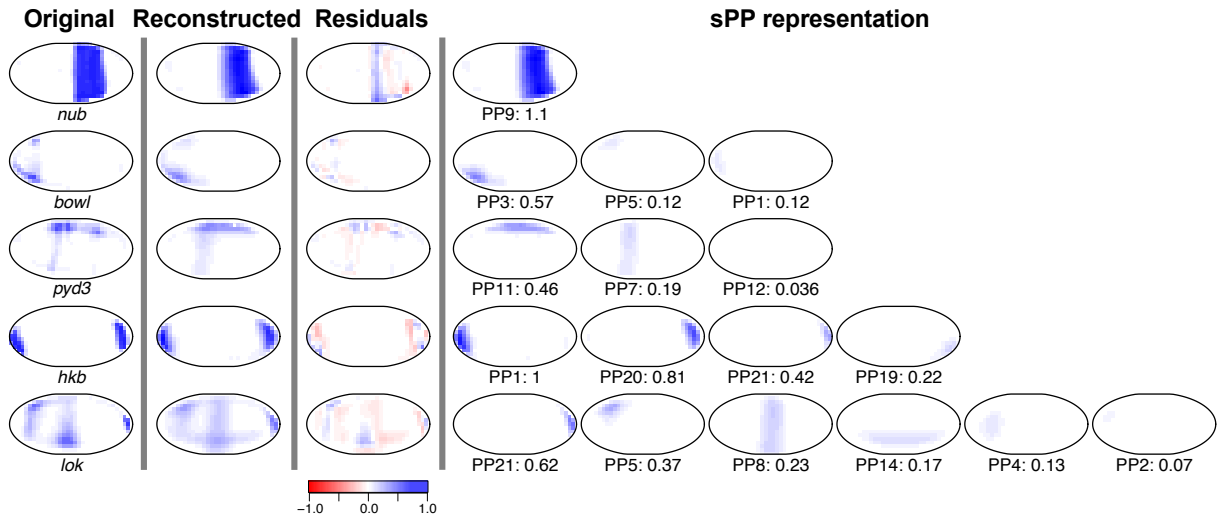

Figure 3.10: Sparse decomposition of spatial gene expression patterns using the LASSO+NLS procedure. Shown are a sample of five gene expression patterns (Original), their reconstructed patterns using the sPP representation (Reconstructed), the difference between the original and the reconstructed patterns (Residuals) and the contributions from the 21 PP (sPP representation).

Chapter 4

Interpreting the learned dictionary – principal patterns (PP)

In Chapter 3, we developed a stability-driven model selection for NMF and derived 21 dictionary atoms, or principal patterns (PP), based on our *Drosophila* gene expression data. In this chapter, we will show that these computationally learned PP are biologically interpretable. First of all, we will link PP to regions of the well-established *Drosophila* pre-organ fate map. Next, we will demonstrate PP's unique biological interpretability by comparing them with dictionary atoms learned from three other unsupervised learning methods: principal component analysis, factor analysis and independent component analysis. Finally, we will use the PP-based representation to predict manual annotations of gene expression. The high accuracy in the prediction task and the interpretability of the top predictors indicate that PP can become a data-driven alternative to the traditional vocabulary-based approach.

4.1 PP and the *Drosophila* fate map

The 21 learned PP divided the *Drosophila* embryo into contiguous regions (Figure 4.1). Each PP is spatially coherent: the intensity is locally continuous and the regions defined by the PP are interconnected. We grouped the 21 PP into four categories: PP1–5: anterior patterns; PP6–9: vertical (gap) segmentation stripes; PP10–16: horizontal ventral-dorsal patterns and; PP17–21: posterior patterns. Furthermore, the PP resemble the pre-tissue and organ regions in the *Drosophila* fate map [35, 36], an experimentally determined functional mapping of spatial regions before availability of gene expression data. In the following paragraph, we will describe in detail how we linked the PP to fate map.

The *Drosophila* fate map is a schematic diagram depicting pre-organ regions of a *Drosophila* embryo. To link the computationally derived PP to the fate map, we first identified a few PP that definitely belong to certain regions of the schematic map and assigned the rest according to their relative positions and shapes. Additionally, we validated our assignments by finding genes with known biological roles using the PP categories described in Chapter

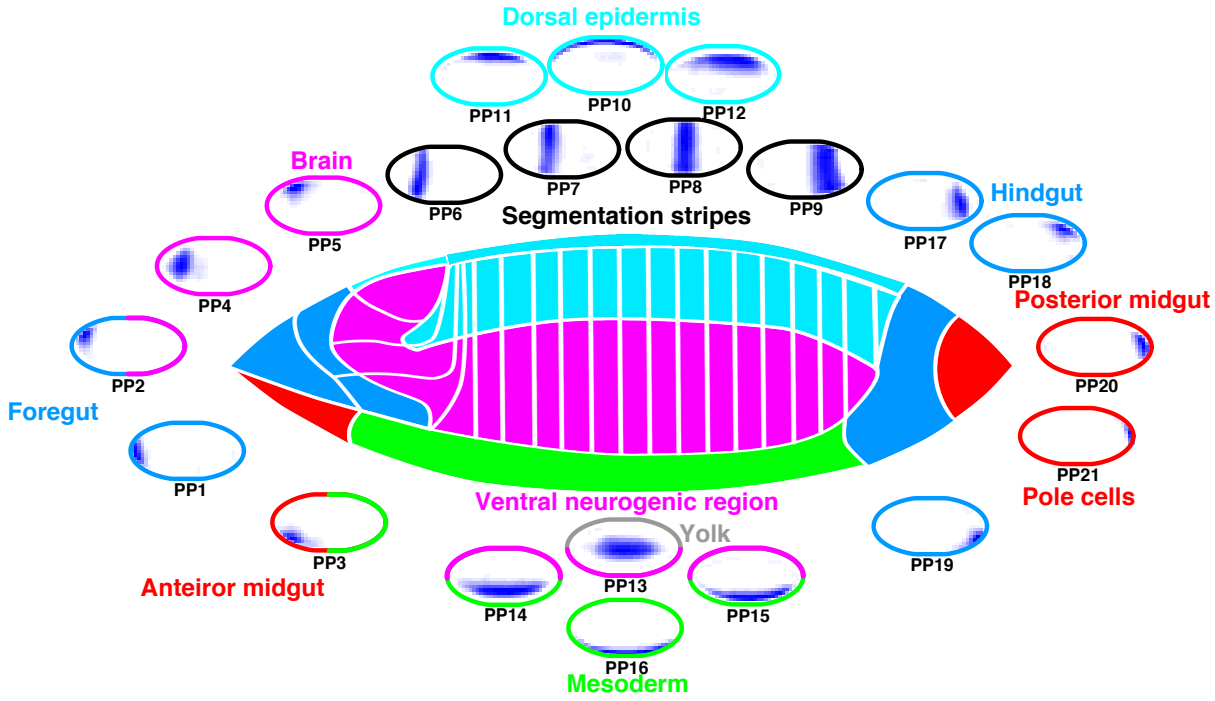

Figure 4.1: The *Drosophila* fate map (center) [35, 36], surrounded by the 21 PP learned by staNMF. The PP are arrayed according to the corresponding regions of the fate map.

5 and later stage annotation data [21]. For example, PP1 can be easily mapped to foregut, PP10 to dorsal epidermis, PP16 to ventral mesoderm and PP21 to pole cells as these PP occupy the four corners of the embryo. Next, for the anterior patterns, we identified PP4 and PP5 from their locations and shapes as the brain region of the fate map. These assignments were further substantiated by a number of nervous system genes associated with the two PP (e.g., *numb*, *oc*, *D* and *Doc1*). PP2 is between the brain and the foregut region and we found genes expressed in PP2 associated with either organ, e.g., *hb* and *tll* in brain and *oc* and *hbn* in foregut. Therefore we labeled it as either brain or foregut. PP3 is most likely to be the anterior midgut or anterior ventral mesoderm regions because it is directly beneath PP1 (foregut) and overlaps with PP16 (mesoderm). Genes expressed in PP3 include known midgut genes (e.g., *egg* and *ry*) and known mesoderm genes (e.g., *croc* and *Mes4*). PP6–9 are vertical segmentation patterns that were hinted in the fate map [49]. For the horizontal patterns, PP11 and PP12 are both above the midline of the embryo and hence can be treated as dorsal epidermis region. The below embryo midline PP14 and PP15 can be either the ventral neurogenic region or mesoderm, and there is evidence supporting that parts of the later central nervous system is derived from the mesoderm [61]. PP13 is most often associated with the ventral neurogenic regions (e.g., *SoxN* and *ind*) and the yolk region of the embryo (not part of the fate map) (e.g., *aay* and *llp4*). For the posterior patterns, PP20

is directly to the left of PP21 (pole cells) and so we labeled it as midgut. This mapping is also supported by the fact that many midgut genes are expressed in PP20, including *sc*, *Bgb*, *esg* and *Moe*. PP17 is labeled as hindgut since it is similar in shape and size to the hindgut region of the fate map. PP18 and PP19 are directly above and below PP17 respectively and so they were labeled as hindgut as well. Moreover, we found hindgut genes such as *Abd-B*, *Mkp3* and *D19A* in PP17, *Doc1*, *ebi* and *dm* in PP18 and *byn*, *apt* and *twi* in PP19, further supporting our mapping of PP17–19 to the fate map.

We found that the PP refined the fate map in the dorsal epidermal region, the ventral neurogenic region, the mesoderm and the hindgut. Some of the refinements are already biologically supported. For example, the vertical stripes are known to be the result of gap, pair-rule, polarity and segmentation genes that eventually establish 14 refined stripes that become morphologically distinguishable in a later stage embryo [49].

4.2 Comparison with factor analysis, PCA and ICA

Can other dictionary learning algorithms recover similar part-based representations as NMF? In this section, we will compare our NMF derived dictionary with those obtained by a sparse Bayesian factor model [24], principal component analysis (PCA) and independent component analysis (ICA) [62]. We will show that only PP recapitulates the underlying biology of cell and tissue fate map.

Recently, a sparse Bayesian factor model was developed to derive patterns from *Drosophila* gene expression data [24] . We applied their algorithm to our images. Since their algorithm involved MCMC computation, we did not perform model selection for the factor model. Instead, we set the number of Bayesian factors (BF) to be $K = 21$ as in NMF for direct comparison (Figure 4.2). For some of the BF, e.g., BF1, 2, 3, 10, 13, 20 and 21, the intensity in the negative region is rather uniform and the BF can be associated with the corresponding PP, e.g., PP1, 2, 3, 11, 16, 20 and 21. However, the biological meaning of the remaining BF is not immediately clear. For example, BF4 splits the mesoderm region into a positive half and a negative half, with other positive and negative regions scattering around the embryo. BF9 seems to be made up from the positive PP8 and the negative PP9. By allowing negative values in the sparse linear coefficients, some of the BF also appeared to be much broader than the PP, e.g., BF5, 8 and 16.

Similarly, we compared our PP with PCA and ICA (Figure 4.3). For PCA, as a consequence of the orthogonality constraint, the derived components show oscillating patterns and are difficult to interpret. For ICA, the negative components in the IC make them less interpretable. We used the R package `FastICA` for the computation of ICA [62] .

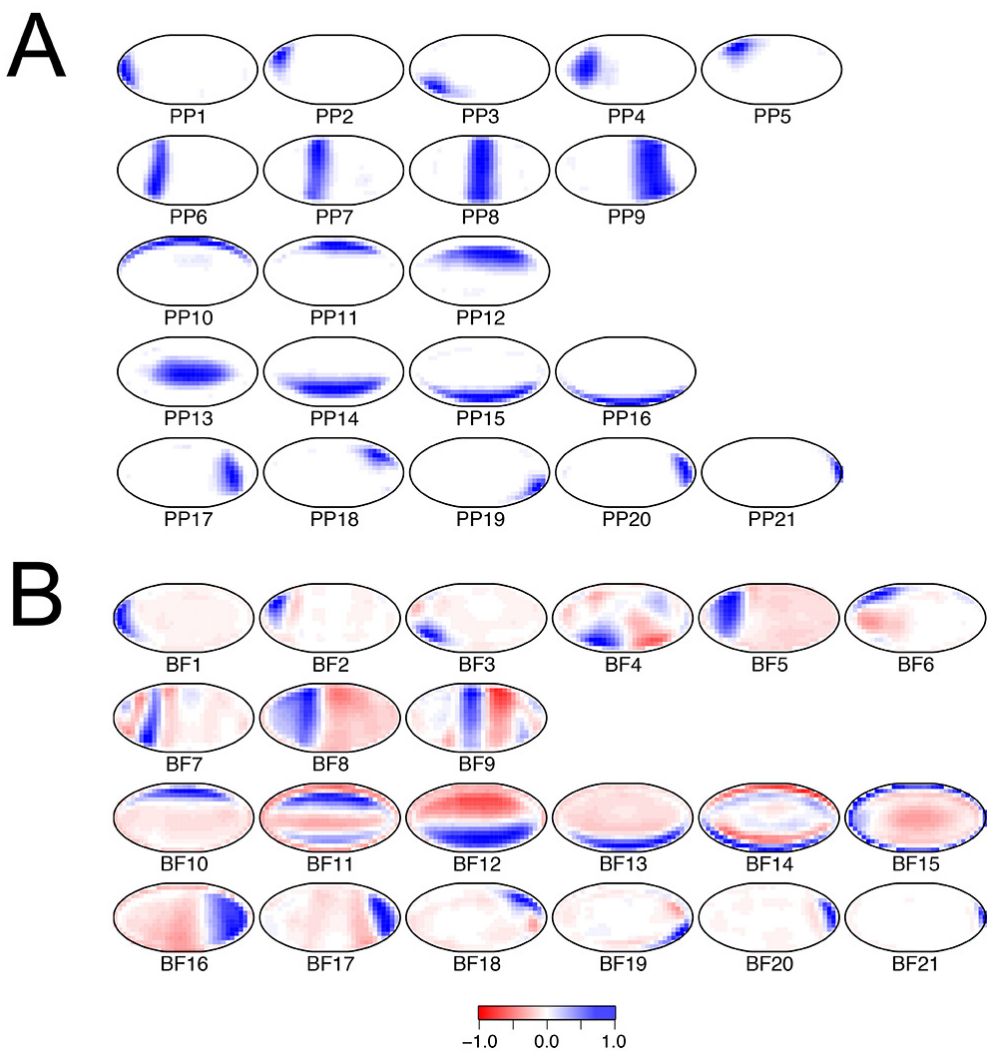

Figure 4.2: A comparison between the 21 principal patterns (PP) and the 21 sparse Bayesian Factors (BF) [24]. (**A**) The 21 learned PP. Every PP was normalized to have maximum intensity equal to one. (**B**) The 21 learned BF. Blue intensity indicates positive value and red indicates negative values. Every BF was normalized to have maximum absolute intensity equal to one.

4.3 PP provide a data-driven alternative to human expert annotations

Traditionally, expert curators annotated BDGP spatial gene expression patterns with a number of controlled vocabulary terms [9, 17, 18, 19, 20, 21]. These terms represent anatomical regions of the developing embryo, similar to the fate-map discussed above. To compare

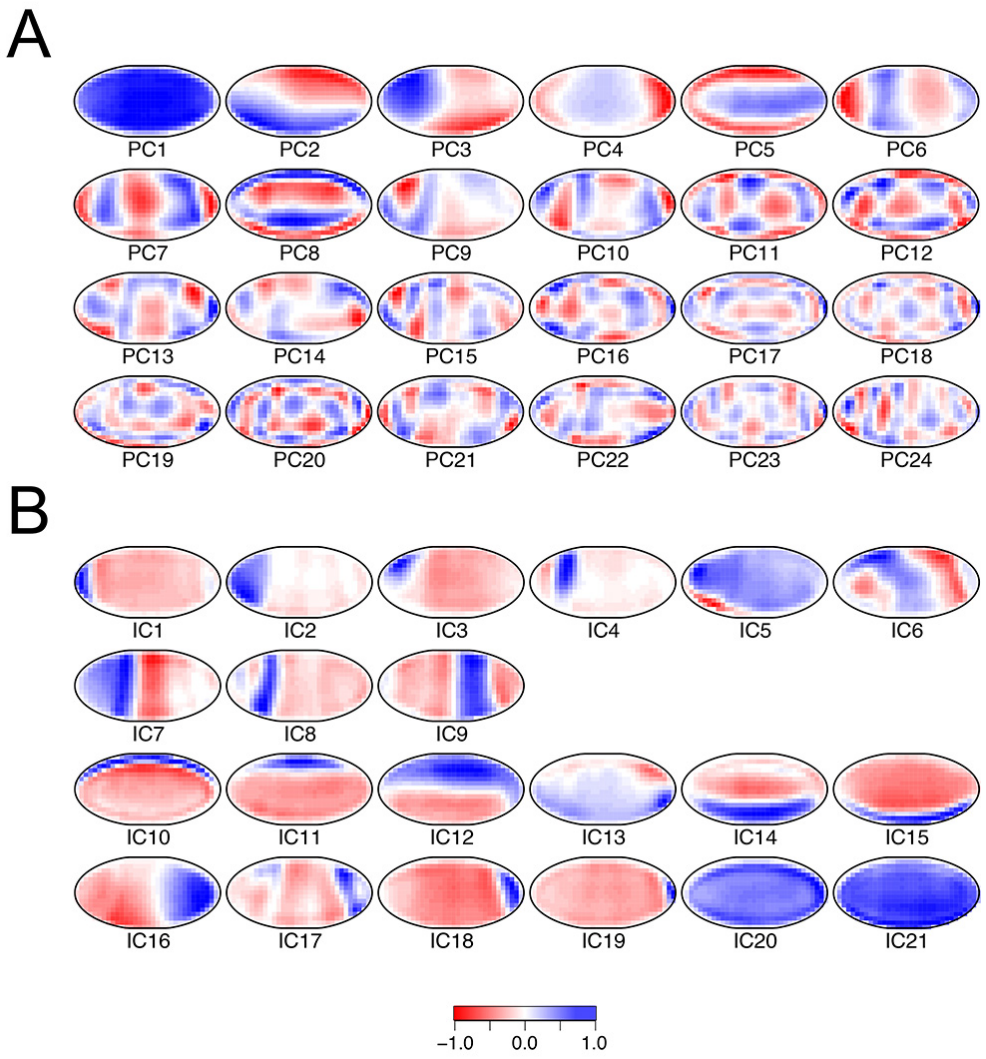

Figure 4.3: Principal component analysis (PCA) and independent component analysis (ICA) for the *Drosophila* gene expression data. (**A**) The top 24 principal components as ranked by the corresponding eigenvalues. (**B**) The 21 independent components.

the 21 learned PP with the anatomical vocabulary, we used the sPP coefficients generated in Chapter 3 as predictors in a supervised learning approach to classify annotation terms. We selected 11 stages 4–6 annotation terms with more than 100 images: ectoderm anlage in statu nascendi (AISN), dorsal ectoderm AISN, procephalic ectoderm AISN, ventral ectoderm AISN, mesoderm AISN, trunk mesoderm AISN, amnioserosa AISN, gap, hindgut AISN, pole cells and visual AISN. For each of the 11 terms, we labeled images annotated this term as "1", the rest as "0". We then fitted an l_1-penalized logistic regression (L1LR) with the sPP coefficients as predictors (see e.g., [63]). To compare with sPP, we also trained

L1LR using the full expression pattern with 405 pixels, and the sparse Bayesian Factor (BF) model by [24] (see previous section).

Specifically, for each annotation term, denote by $L[i]$ the label of the i-th image for $i = 1, ..., 1640$: $L[i] = 1$ if the gene corresponding to the image was labeled as expressed in this term and $L[i] = 0$ otherwise. To predict the label vector L, we fitted L1LR using three different covariate sets: (1) the 405 pixels for the pixel-based representation, (2) the 21 sPP coefficients based on the LASSO+NLS procedure, and (3) the 21 sparse Bayesian Factor (BF) coefficients. For each annotation term, the observations in each class are weighted by the reciprocal of the corresponding class size so that the two classes are of the same importance. A 10-fold cross-validation was performed and the l_1-penalization parameter was chosen such that it was the largest among all parameters whose cross-validation Area Under the ROC Curve (AUC) was within one standard error of the maximum AUC. We used the R package `glmnet` [60] for computation.

Prediction performance and model complexity

The prediction performance of the three representations is very similar, as measured by the cross-validation AUC (Figure 4.4*A*). On average, the AUC value for the sPP representation is 0.772, as compared to 0.787 for the pixel-based representation and 0.767 for the BF representation. Taking into account the standard error of the AUC for each annotation term, none of the three methods significantly outperforms the others. In terms of model complexity, on average 17 predictors are selected for the pixel-based L1LR, 7 for our sPP-based approach, and 8 for the BF-based model (Figure 4.4*B*).

Stability analysis of the selected predictors

We studied the stability of the selected predictor sets for the three representations. To interpret the selected predictors, stability is the minimal requirement. For each annotation term and representation, we measured instability using the Jaccard distance between the supports of two L1LR coefficients, averaged over all 45 coefficient pairs in the 10-fold cross-validation. The higher the Jaccard distance, the more unstable the support of the L1LR coefficients. Our results indicated that the selected L1LR model for the sPP representation is most stable among the three representations, except for two terms: "dorsal ectoderm" and "hindgut", for which the BF approach is slightly better (Figure 4.4*C*). The pixel-based approach selects highly unstable predictor sets.

L1LR coefficients for PP and BF

Next, we examined the L1LR coefficients for PP and BF representations (Figure 4.5 and 4.6). For all 11 terms, the PP L1LR coefficients are sparse and the largest L1LR coefficients are always positive. Furthermore, the top L1LR coefficient – the largest L1LR coefficient in magnitude – is much larger than the second largest L1LR coefficient in magnitude. These

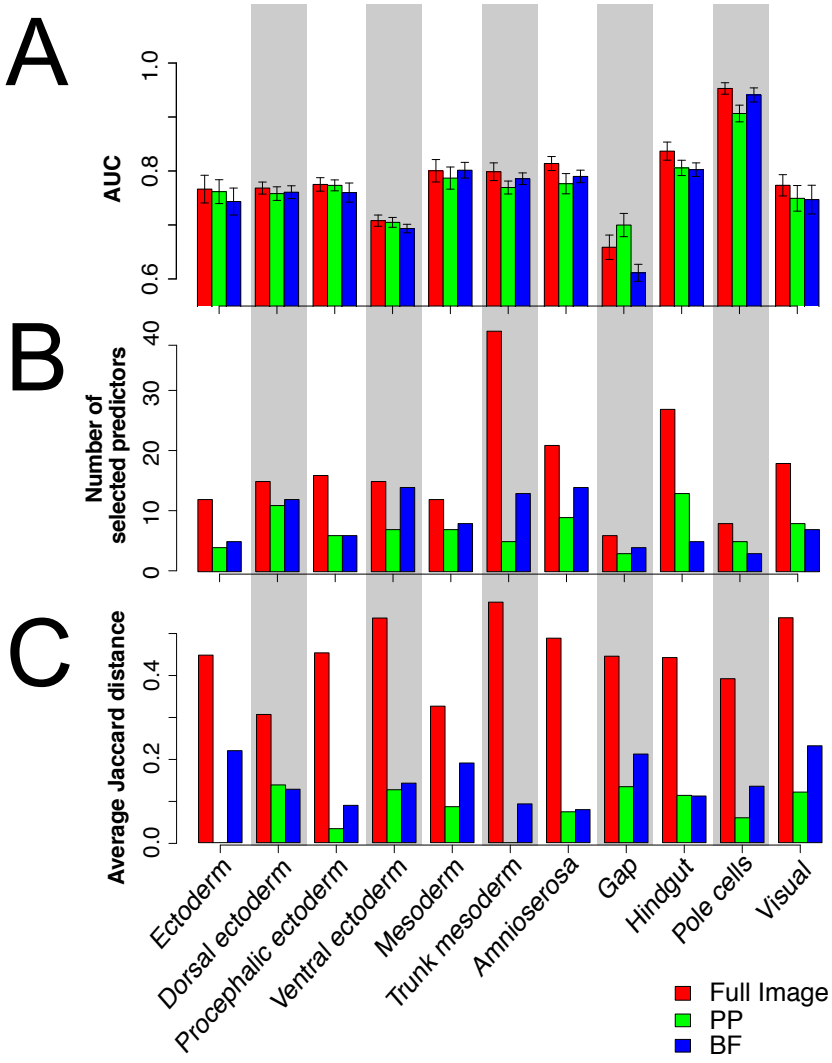

Figure 4.4: Predicting annotation terms based on 405 image pixels, the sPP and the BF (sparse Bayesian Factor) representations. (A) Prediction accuracy as evaluated by the AUC value. Data are expressed as mean ± SEM. (B) Number of selected predictors in the optimal model. (C) Stability analysis of the set of selected L1LR predictors for three representations.

facts indicate that the top L1LR PP – the PP that corresponds to the top L1LR coefficient – is the dominating factor in determining whether a gene expression pattern is labeled with an annotation term. This also gives motivation to consider only the top L1LR PP when associating PP with annotation terms in the next subsection. For the BF representation, the L1LR coefficients are also sparse. But unlike the PP case, the largest L1LR coefficients are not always positive (e.g., "ventral ectoderm" and "gap"). In addition, some of the largest L1LR coefficients are much closer to the second largest L1LR coefficients than in the PP

case: e.g., the terms "mesoderm" and "visual". This is because the positive regions of some BF have a significant amount of overlap, e.g., BF12 versus BF13, and BF5 versus BF6. The overlap between PP is much smaller.

Visualizing L1LR coefficients

For the pixel-based model, we created a visualization of the 405 predictors for each annotation term by plotting the L1LR coefficient values as pixels in our elliptic embryo shape. To compare with this visualization, we selected the (top L1LR) PP and BF corresponding to the largest L1LR coefficients for their respective L1LR models. The pixel-based predictors consist of scattered points and the top L1LR BF contains negative values, both of which are difficult to interpret. In contrast, the top L1LR PP consistently showed the annotation term exactly as a curator would annotate the gene expression.

For some annotation terms, the positive predictors for the pixel based model, the top L1LR PP and the positive part of the top L1LR BF overlapped with the regions in the embryo described by the controlled vocabulary terms. For example, for the annotation terms such as "dorsal ectoderm", "mesoderm", "trunk mesoderm" and "pole cells", the top PP corresponds to the areas of the embryo that can be easily recognized as those anlagen. On the other hand, the selected predictors for the pixel-based representation are predominantly isolated pixels at locations associated with the specific annotation term. Of the 11 annotation terms, all of the top L1LR PP but only nine of top L1LR BF-based components have positive fitted L1LR coefficients. For some of the nine terms with positive association for both PP and BF (e.g., "dorsal ectoderm", "mesoderm" and "pole cells"), the top L1LR PP and the positive part of the top L1LR BF have similar shapes and sizes. For other terms such as "hindgut" and "visual", the positive part of the top L1LR BF pattern appears to be much broader than the top L1LR PP.

Discussion

The PP-based representation of gene expression patterns provides a data driven alternative to a synthetic vocabulary with a manually generated curation approach. Developing a controlled vocabulary requires prior biological knowledge and visible references and computational annotation demands a sizable training dataset, usually hand curated. In our studied dataset of early undifferentiated *Drosophila* embryos, the lack of visible morphological reference features has introduced inconsistencies and errors and does not capture the full richness of the dataset [27]. The early stage annotation dataset is an imperfect gold standard for objective evaluation but still proved valuable for evaluating accuracy and validation of capturing the underlying biology. We identified a PP for every organ specific annotation term in addition to PP not represented by annotations (e.g., segmentation patterns). Using sPP coefficients, spatial expression patterns are easily combined with other data types to facilitate biological modeling efforts.

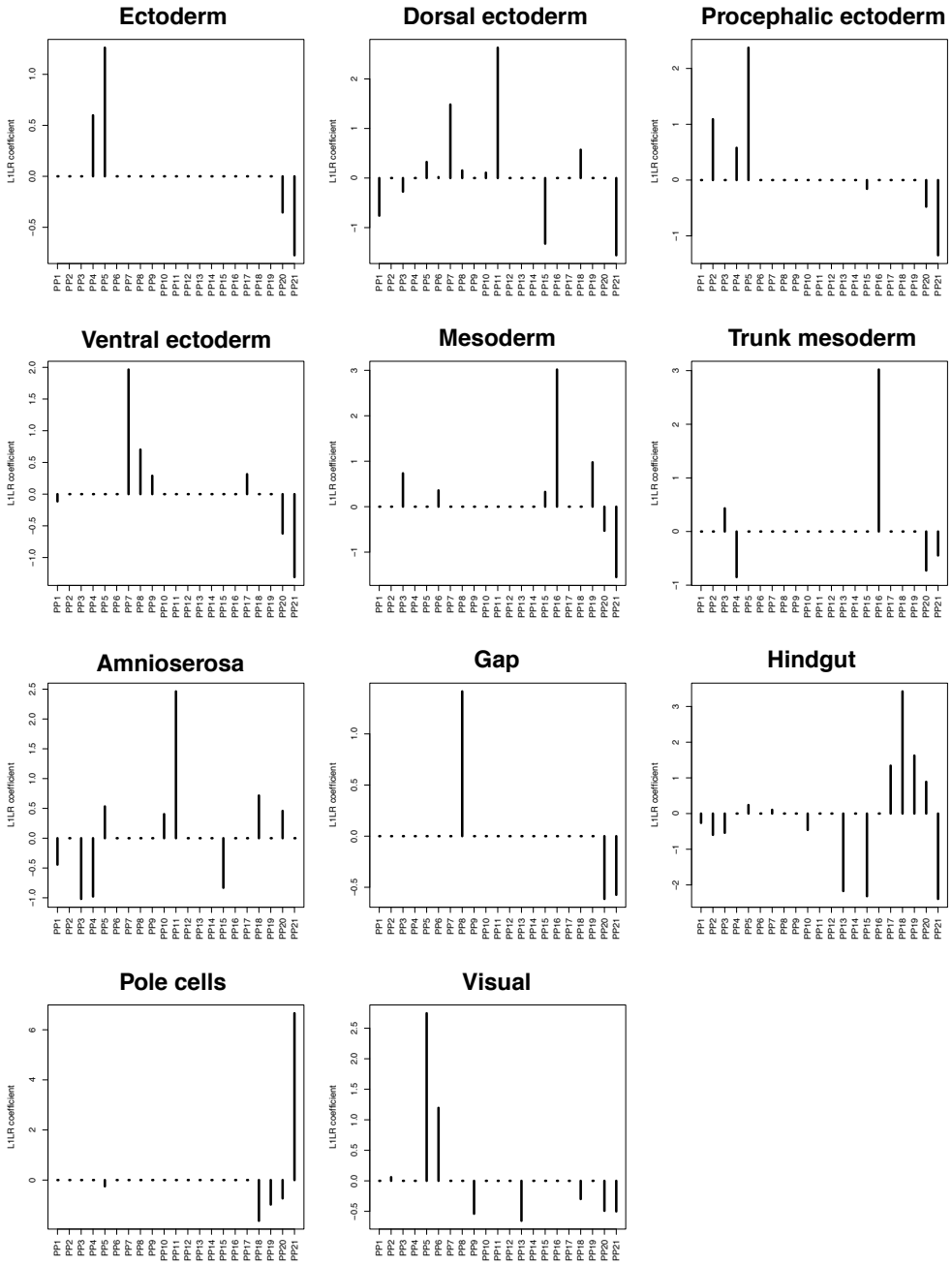

Figure 4.5: L1LR coefficients for annotation prediction using the sPP representation. For all 11 terms, the L1LR coefficients are sparse and the largest L1LR coefficients are always positive. Furthermore, the top L1LR coefficient is much larger than the second largest L1LR coefficient in magnitude. These facts indicate that the top L1LR PP is the dominating factor in determining whether a gene expression pattern is labeled with an annotation term.

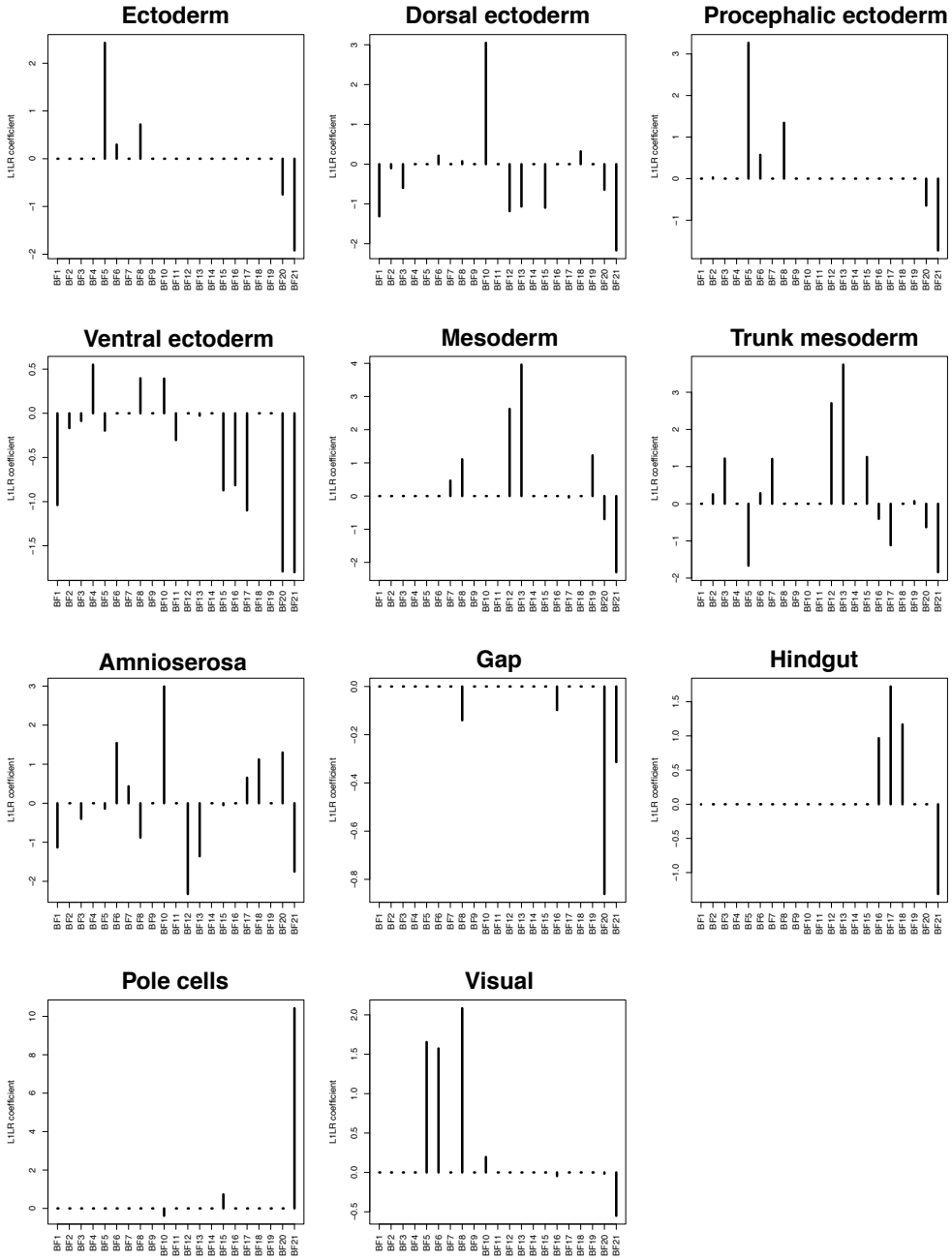

Figure 4.6: L1LR coefficients for annotation prediction using the BF representation. The L1LR coefficients are sparse. But unlike the PP case in Figure 4.5, the largest L1LR coefficients are not always positive ("ventral ectoderm" and "gap"). In addition, some of the largest L1LR coefficients are much closer to the second largest L1LR coefficients than in the PP case: e.g., the terms "mesoderm" and "visual".

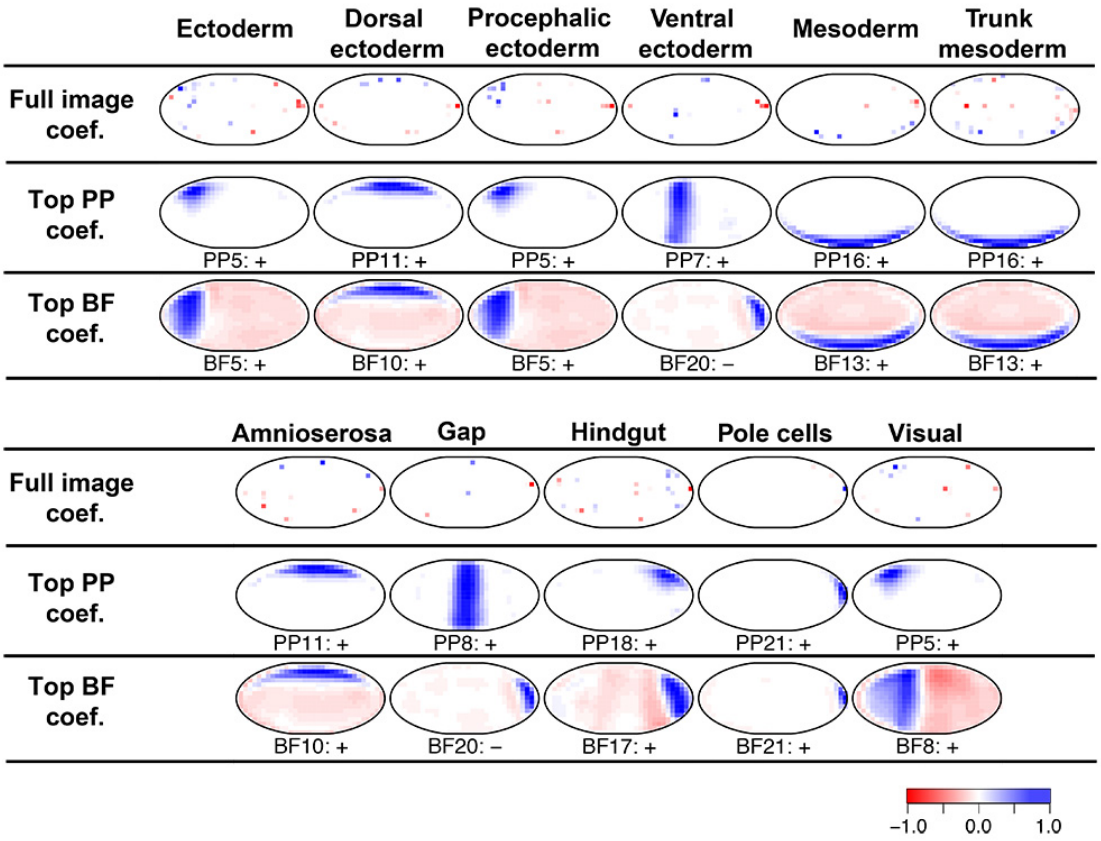

Figure 4.7: Interpretability of the L1LR under the pixel-based, the sPP and the BF representations. The pixel-based full image representation: all L1LR coefficients are shown as pixel values within the embryo; the sPP-based and the BF-based representations: only the top L1LR PP or BF that corresponds to the largest L1LR coefficient is shown. The scale goes from -1 to 1 and is color coded respectively from red to blue. For the PP and BF, "+" indicates that the largest L1LR coefficient is positive, and "-" indicates that the largest L1LR coefficient is negative.

Chapter 5

Functional categorization of genes with PP

We have demonstrated that the staNMF-derived PP are biologically meaningful as they correspond to pre-organ regions in the previously established *Drosophila* fate map. In this chapter, we will utilize the sPP coefficients, obtained through the LASSO+NLS procedure described in Chapter 3, to systematically associate genes to 21 categories. We also investigate the fraction of genes shared by a pair of categories and reveal a link between anterior and posterior embryonic regions. Furthermore, we will relate PP to later stage organ systems, confirming our earlier mapping of PP to the fate map.

5.1 PP associated gene functions

We defined the term "function" by the experimentally generated fate map that describes the locations of larval/adult progenitor cells in the blastoderm. These cells give rise to particular tissues and organs during development. For $k = 1, ..., 21$, we defined the k-th sPP coefficient of a gene to be the maximum k-th sPP coefficient among all the replicate patterns of the same gene. We assigned a gene to *PP category* k if the k-th sPP coefficients of the gene exceeded 0.1. The number of genes in each of the 21 PP categories is, on average, 300 genes ranging from 184 to 395. PP categories 6–9, contain fewer, on average, 223 genes (Figure 5.1 right). In addition, we also found a significant presence of previously uncharacterized computed genes (CG) in all PP categories: the average percentage of CG per PP category is 23.4%.

To directly relate genes to each other, we created a heatmap visualization of the sPP coefficients for 667 genes that belong to at least one PP category. We ordered the genes by first associating each of them to the PP with the maximum sPP coefficient, and then performing a hierarchical clustering of the genes assigned to the same PP (Figure 5.1 left). A surprisingly large fraction of genes (17.8%) exhibit their strongest expression in PP21 (pole cells) and have limited expression in other PP. We found that only 5.8% of the 156 transcription

factors are among these PP21 specific genes, confirming previous results [21]. 4.5% of the 667 genes have their strongest expression in segmentation patterns PP6–9, suggesting that only a small number of genes are dedicated to segmentation. Furthermore, 93.3% of these genes have been characterized, implying that we know most segmentation genes. We found genes with known roles in foregut development (*croc*, *hkb* and *kni*) associated with PP1, segmentation specific genes (*Dfd*, *kn*, *Kr* and *tsh*) associated with PP6–9, genes essential for mesoderm/ectoderm development (*mes*2, *sna* and *sog*) associated with PP15, genes essential for pole-cell formation associated with posterior PP21 (*lok*, *pgc* and *rdx*) as well as previously uncharacterized genes such as *CG1663*, *CG8289*, *CG9514* and *CG10479* in these PP categories (Figure 5.2).

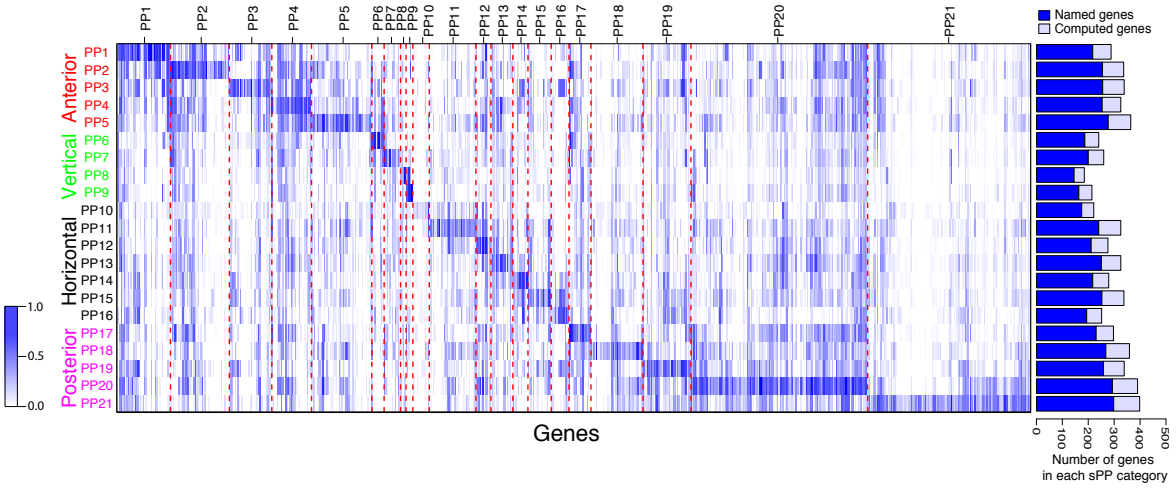

Figure 5.1: PP-based gene categorization. Left heatmap: PP expression profile of genes. Each column corresponds to the sPP coefficients of one gene. Between the red dashed lines are the genes with the strongest expression in the same PP. Right barplot: numbers of named and uncharacterized computed genes (CG) in each PP category.

5.2 Relationships between spatial regions

Next, we investigated the relationship between the PP that span the anterior-posterior axis, i.e. PP1–9, PP17–21. We plotted the fraction of common genes in a pair of PP categories, defined as the Jaccard distance between the two categories, in relation to the pairwise PP centroid distance (Figure 5.3). Our results show that when the PP distance is small, the fraction of common genes is high. However, after the initial decrease, the fraction of common genes increases as the PP distance increases. An example is the set of genes (49% or 227) shared between the distant PP2 and PP18 that map to anterior foregut/brain and posterior

hindgut (Figure 5.3). These genes include known foregut and hindgut development genes such as *Alh*, *Blimp-1*, *Btk29A*, *dm*, *Mkp3* and *rpr*. This finding substantiates the previously identified common origins and gene expression signatures of foregut and hindgut that were based on manual annotations [36, 21]. Similarly, 229 genes (52%) are shared between PP3 (anterior midgut/mesoderm) and PP19 (hindgut), including known midgut and hindgut genes, *ry*, *Ect4*, *Sdc*, *Pcl*, *larp* and *emc*, suggesting a more general link between the anterior and posterior patterns.

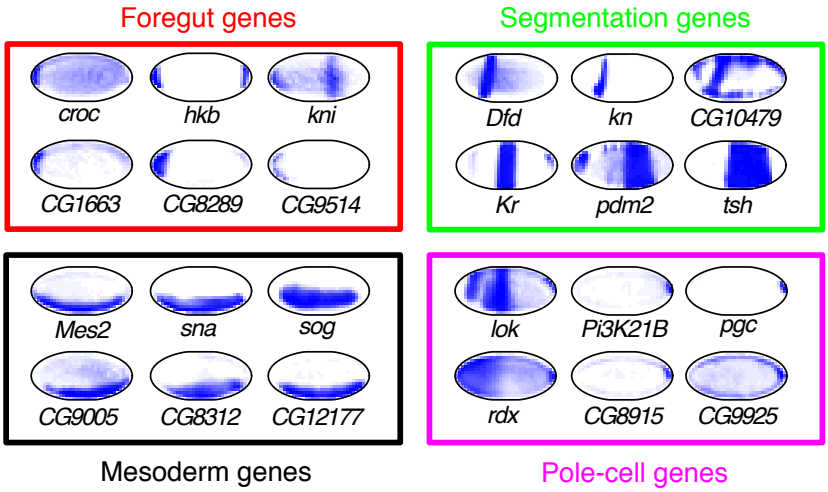

Figure 5.2: Known genes and uncharacterized computed genes (CG) were found in the associated PP categories.

5.3 Linking PP and future organ systems

Using our manual annotations, we related gene expression during late organ system (OS) formation to the PP derived from early embryonic gene expression data. Our annotation data contains OS label information for the next four developmental stages, i.e., stages 7–8, 9–10, 11–12, 13–16. In our analysis, we selected the following eight OS: visual primordia system (VisualPr), central nervous system (CNS), ectoderm or epidermis (Ect/Epi), foregut, midgut, hindgut, mesoderm/muscle (Meso/Muscle) and pole cells. For each stage and OS combination, we compared the sPP coefficients of genes annotated in the OS to the remaining genes using the Mann-Whitney test and plotted the negative logarithm of the p-values (Figure 5.4). We found that genes with high expression intensity in PP5, PP10, PP1, PP20, PP18, PP15-16, PP21 at stage 4-6 are expressed in the tissue corresponding to their fate map position, VisualPr or CNS, Ect/Epi, foregut, midgut, hindgut, mesoderm, pole-cell respectively. The early mesoderm genes (PP14–16) become expressed in the CNS starting at stage 9 (*trx*, *sna*, *Traf4* and *Caf1*). Early mesoderm genes with function during CNS development

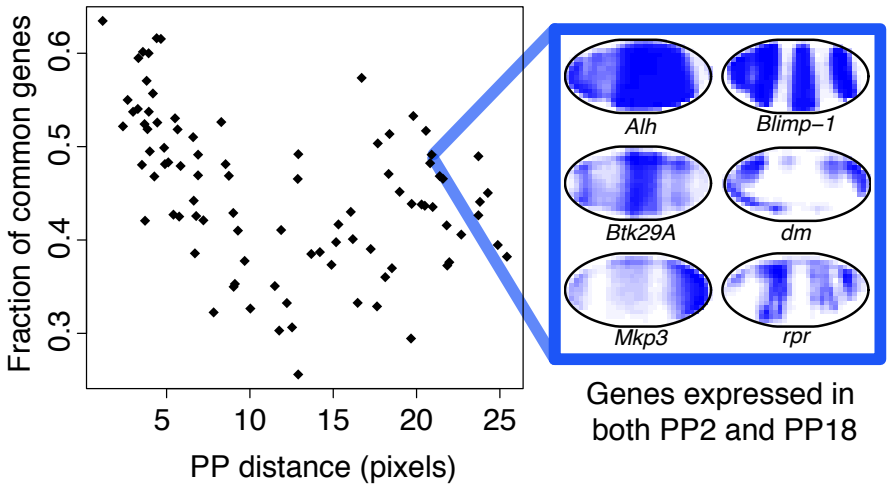

Figure 5.3: The relationship between the fraction of common genes in a pair of PP categories and the centroid distance of the two PP, for PP1–9, PP17–21. Each dot in the plot corresponds to a PP pair. Shown also are six genes expressed in both PP2 (brain/foregut) and PP18 (hindgut), a pair of distant PP.

have been shown before [61], but here we demonstrate a systematic secondary function of mesoderm specific genes, including previously uncharacterized genes (e.g., *CG11247*). Genes in PP from all fate regions appear in the midgut genes at later stages, probably due to its endodermal origin. Finally, genes originally mapped to the ventral epidermis (PP10) are strongly present in the foregut at stages 9-12. In contrast, as expected, OS with no mappings to the fate map, show no clear bias at later stages. Thus, with molecular data, here we show for the first time a systematic relationship between the fate map and gene expression during organ system specification, but not during differentiation.

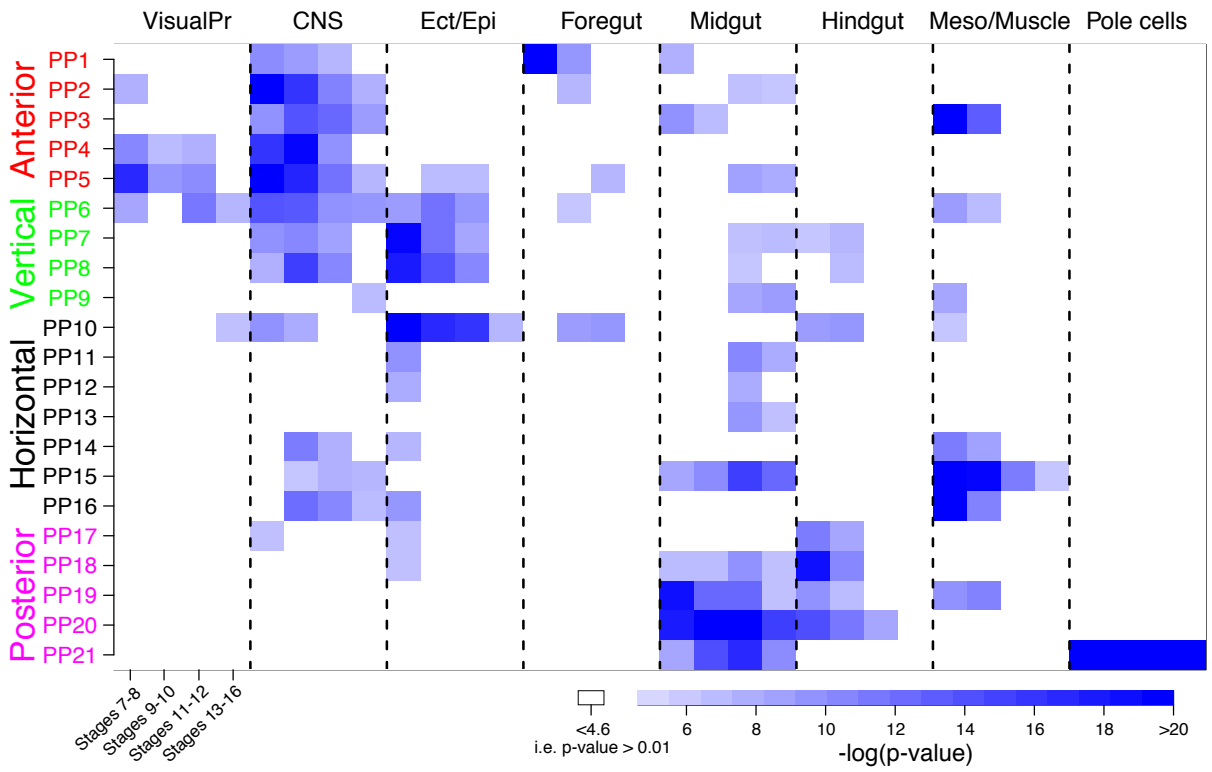

Figure 5.4: Relating gene expression during later organ system (OS) formation to the early stage PP. For each OS and stage range combination, we split the k-th sPP coefficients into two groups: Group 1 contains the coefficients corresponding to genes that were labeled as expressed in the OS and at the stage range, and Group 0 contains the coefficients for the remaining genes. We then performed a Mann-Whitney test for the above two samples. Shown is the heatmap for the negative logarithm of the p-values.

Chapter 6

Spatially local correlation networks

Associations between two genes are routinely described by their correlation to each other [64]. In terms of spatial relationships, positive gene interactions exhibit spatial overlap whereas repressive gene interactions exhibit spatial exclusivity. In the past, attempts to find gene networks from spatial gene expression datasets correlated patterns over the whole embryo [27, 30]. However, spatial gene expression patterns are composed of sub-patterns driven by distinct enhancer elements [33, 34]. Gene interactions are thus context sensitive and not necessarily spatially uniform. In this chapter, for the first time, we will introduce a learned local aspect – the use of PP – to identify gene interactions contingent on local expression context. Spatially local correlation networks (SLCN) are constructed for 156 transcription factors (TF) from our expression data. Our approach is simple, computationally efficient, and intuitive as it mimics the way that a human curator would search for potential interacting patterns. We further test our SLCN using the well-known gap-gene network and are able to recover 10 out of 11 links.

6.1 PP-based correlation network construction

The *Drosophila* gap gene network has been studied for decades [49, 50, 51]. It controls embryonic patterning by regulating the genes required to establish the anterior/posterior segmentation stripes and is primarily driven by well studied activating and repressive interactions between eight TF. To reconstruct this network solely from our expression data of 156 spatially restricted TF, we selected six PP (PP6–9, PP17 and PP20) corresponding to the domains of the gap gene network. We called the six PP *gap-PP*. For each gap-PP, we identified its directly adjacent PP by visual inspection (Table 6.1).

For each of the six gap-PP, we found all TF expression patterns in the category of the gap-PP, or its directly adjacent PP, with sPP coefficient greater than threshold 0.1. This excluded TF with low or no expression in the gap-PP and its nearby regions and hence reduced the possibility of spurious correlations. Denote this set of patterns by T. We then computed the weighted correlations for the expression patterns in T with the l_1-normalized

	Adjacent PP
PP6	PP4, PP7
PP7	PP6, PP8
PP8	PP7, PP9
PP9	PP8, PP17
PP17	PP9, PP20
PP20	PP17, PP21

Table 6.1: The adjacent PP for the six gap-PP.

PP intensity as the weight vector. Specifically, let $\mathbf{u} \in \mathbb{R}^d$ be a nonnegative vector whose entries sum up to one and $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ represent two data vectors, e.g., two gene expression patterns in T. The local correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$ with weight $\mathbf{u}$ is defined as:

$$\text{cor}_{\mathbf{u}}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\text{cov}_{\mathbf{u}}(\mathbf{x}_1, \mathbf{x}_2)}{\text{var}_{\mathbf{u}}(\mathbf{x}_1)^{1/2}\text{var}_{\mathbf{u}}(\mathbf{x}_2)^{1/2}},$$

where

$$\text{var}_{\mathbf{u}}(\mathbf{x}_1) = \sum_{j=1}^{d} \mathbf{u}[j](\mathbf{x}_1[j] - \mathbf{x}_1^T\mathbf{u})^2, \text{ and}$$

$$\text{cov}_{\mathbf{u}}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{d} \mathbf{u}[j](\mathbf{x}_1[j] - \mathbf{x}_1^T\mathbf{u})(\mathbf{x}_2[j] - \mathbf{x}_2^T\mathbf{u}).$$

Note that when the $\mathbf{u}[j] = 1/d$ for all $1 \leq j \leq d$, the above correlation is the same as the sample correlation between data vectors $\mathbf{x}_1$ and $\mathbf{x}_2$.

As mentioned, many genes had multiple replicate images. For a pair of genes, we defined the local correlation of the two genes to be the local correlation with the maximum magnitude between replicate images of one gene and replicate images of the other. For simplicity, we called this correlation the maximum correlation, although we note that it can be the most positive or the most negative correlation. By computing this maximum correlation, we stated that two genes were highly correlated if any of the replicates of the two genes were highly correlated. Spatial expression patterns for some genes change rapidly within the stage range considered in this thesis. For example, significant differences in gene expression were observed for the replicate expression patterns of *hb* and *kni* (Figure 2.3). Using maximum local correlation can therefore help to identify those highly variable genes that were likely to interact at some point in the developmental timeline.

For each gap-PP, we computed the local correlation for all pairs of genes in the gene set T defined earlier. The distribution of the correlations was bimodal, with one peak corresponding to positive correlations and the other to negative correlations (Figure 6.1). This is due to the way we defined local correlation of two genes, which excluded image pairs that had close-to-zero correlations. To construct the local network for each gap-PP, we set

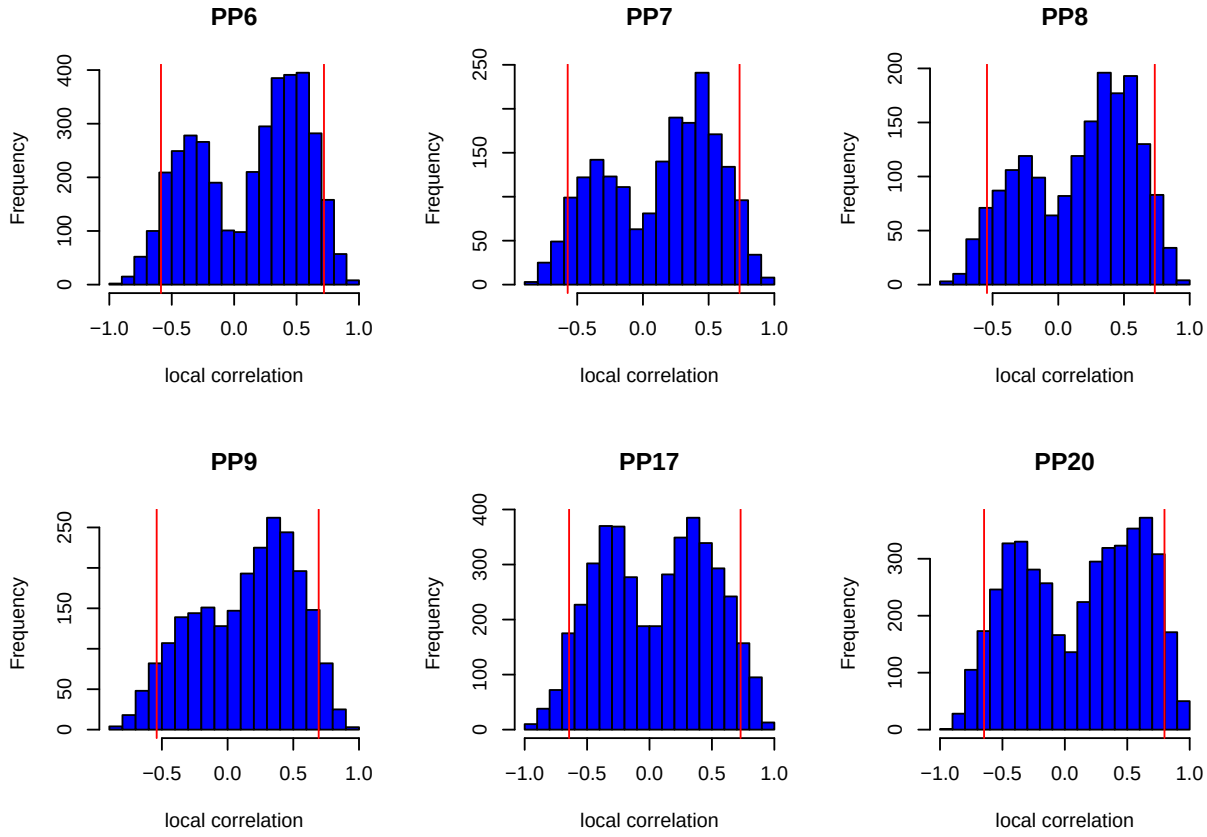

Figure 6.1: Histograms of local correlations for the six gap-PP. The two red vertical lines in each histogram indicate the lower and upper five percentiles of the local correlations, respectively.

a positive edge between two genes if their local correlation is above the upper five percentile of all local correlations for the PP and set a negative edge between two genes if their local correlation is below the lower five percentile. We call the resulting network *spatially local correlation network*, or SLCN.

6.2 Evaluation of SLCN with the gap gene network

We evaluated our SLCN construction by comparing interactions found in the six SLCN to known regulatory interactions of selected trunk and terminal gap genes, $giant(gt)$, $hunchback(hb)$, $knirps(kni)$, $Krüppel(Kr)$, $huckebein(hkb)$ and $tailless(tll)$. We compared the subnetworks of the SLCN containing only the six genes (Figure 6.2A) to a schematic network diagram (Figure 6.2B), as originally depicted in [51]. While the diagram indicates that some gene

interactions are contingent on spatial position, it does not provide precise locations of the interactions. To compare with our networks, we devised a method to match the links in the diagram to our SLCN. For each gap gene, we first created a linearly ordered PP representation by placing the six gap-PP anterior to posterior and associating a gap-PP to the gene if the sPP coefficient for the gap-PP exceeded a threshold of 0.1 (Figure 6.2*C*). The gap-PP associated with each gap gene were then merged into one or more connected PP groups. Based on its relative location in the diagram, we then matched each gene node in the schematic diagram to a connected PP group for the same gene. We considered an interaction between two gene nodes in the schematic network diagram as successfully identified by our method if the same interaction exists in any SLCN associated with the overlapping PP in the connected PP groups of the two gene nodes.

For example, the diagram (Figure 6.2*B*) depicts a repressive link between the anterior component of *gt* (i.e. gt_1) and Kr. Using our linearly ordered PP representation, we found the connected PP groups for gt_1 and Kr are PP6/7 and PP8, respectively (Figure 6.2*C*). We searched for the *gt*-Kr interaction only in the SLCN of PP7 and PP8, since PP6 and PP8 do not overlap. In both networks, we found a repressive interaction (or negative correlation). Hence we considered the anterior *gt*-Kr link of the schematic gap gene network diagram as being identified with our model. See Table 6.2 for the validation of the remaining links.

For the six gap genes, our SLCN reconstruction identified 14 interactions (Figure 6.2*A*). Eight out of 11 links in the gap gene network diagram have a one-to-one mapping with eight of the 14 SLCN interactions. In addition, the two *gt*-Kr links in the gap gene network (Link 1 and 5 in Figure 6.2*B*) are found in the SLCN of PP7–9 (Links 2, 4 and 7 in Figure 6.2*A*). The remaining *kni*-gt_2 link (Link 6 in Figure 6.2*B*) has no corresponding link in the SLCN. Therefore, our SLCN recovered 10 out of 11 interactions in the gap gene network. Three of the 14 SLCN links do not correspond to any interactions in the network diagram. In PP6, we found a repression link between *gt* and *kni* (Link 1 in Figure 6.2*A*). Gene expression images of *gt* and *kni* revealed a clear complementary pattern towards the anterior end with a negative local correlation of -0.720 in PP6. In the PP17 SLCN, an activation link between *kni* and *gt* was identified (Link 11 in Figure 6.2*A*). Since our images covered an interval of around 1.5 hours, the posterior part of *kni* expression pattern at the early developmental stages 4-6 might have been aligned to the *gt* gene posterior end at a later time point. Experiments are needed to confirm or refute these predicted links. Finally, although not described in [51], the predicted *hb*–*tll* activation link in PP9 (Link 6 in Figure 6.2*A*) is supported by [65].

6.3 Correlating genes on the whole embryo

We compared our PP-based local network results to those obtained by correlating the expression patterns over the whole embryo, or global correlation analysis. Similar to local correlation, we defined the global correlation of the two TF as the largest correlation between the replicates of the two. Next, we specified a cutoff value for the global correlation in order to form network links. We first combined the six PP-based SLCN into a single network

Link in gap gene network	G1	G1 PP	G2	G2 PP	Overlapping PP	Link(s) in SLCN
1	gt_1	PP6,7	Kr	PP8	PP7,8	2,4
2	hb_1	PP6-8	Kr	PP8	PP7,8	3
3	hb_1	PP6-8	kni	PP8,9	PP7-9	5
4	kni	PP8,9	Kr	PP8	PP8,9	8
5	gt_2	PP9	Kr	PP8	PP8,9	4,7
6	gt_2	PP9	kni	PP8,9	PP8,9	No link
7	hb_2	PP17,20	kni	PP8,9	PP9,17	12
8	gt_2	PP9	hb_2	PP17,20	PP9,17	9
9	kni	PP8,9	tll	PP17,20	PP9,17	13
10	gt_2	PP9	tll	PP17,20	PP9,17	10
11	hb_2	PP17,20	hkb	PP20	PP17,20	14

Table 6.2: Validating the SLCN with the gap gene network. Link in gap gene network: link number in the schematic gap gene network (Figure 6.2B). G1 and G2: gene nodes in the schematic gap gene network. G1 PP and G2 PP: the connected PP group in the linearly ordered PP representation that correspond to G1 and G2 respectively (Figure 6.2C). Overlapping PP: the overlapping PP of G1 PP and G2 PP. Link(s) in the SLCN: the link(s) in the predicted SLCN (Figure 6.2A) that correspond to a link in the schematic gap gene network. Out of 11 links in the schematic gap gene network, there is one (i.e. Link 6) that has no corresponding link in the SLCN. There are three links out of 14 in the SLCN that have no corresponding links in the gap gene network diagram.

such that two TF share a link in the new network if they share a link in at least one of the six SLCN, regardless of the sign of the link. For fair comparison between the global and local approaches, the cutoff values for the global correlation network was chosen such that (1) the resulting network has the same number of links as in the previous combined network and, (2) the number of positive links is the same as the number of negative links. We converted the original schematic gap gene network to the "global version" without the spatial information accordingly: two gap genes share a link if they share a link in the schematic gap gene network diagram regardless of the location of the interaction (Figure 6.3C). Only three links out of nine links in the global version of the gap gene network were recovered (Figure 6.3D). An analysis of the relationship between the local and global correlations indicated that, while for some gene-gene interactions global correlation is positively correlated with local correlation, many others have negative correlations (Figure 6.3A). For example, gt and hb are known to be mutual repressors of one another towards the posterior end of the embryo. The global correlation was unable to detect this relationship whereas the local correlation succeeded (Figure 6.3B).

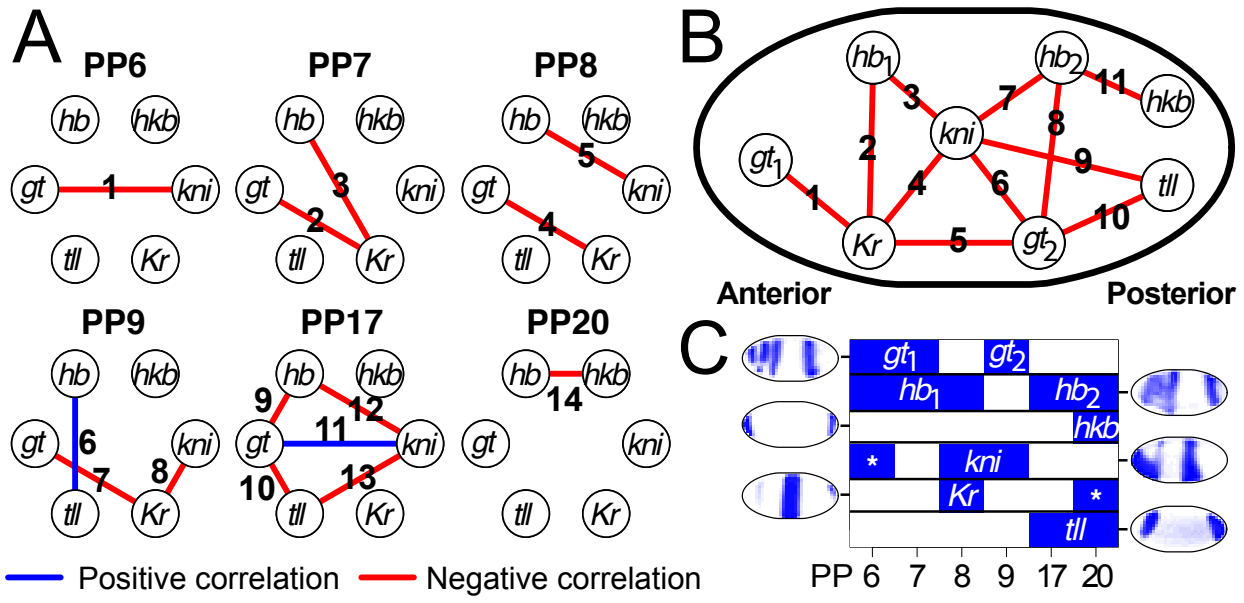

Figure 6.2: Modeling and validation of the Drosophila gap gene network with spatially local correlation networks (SLCN). (A) The SLCN for six gap genes. For each of the six gap-PP, shown is the sub-network of the SLCN that contains the six gap-genes. Links are numbered from 1 to 14. (B) The gap gene network diagram depicting repressive interactions of six genes [51]. Links are numbered from 1 to 11 and multiple occurrence of the same gene are subscripted by numbers (e.g., hb_1 and hb_2). The directions of the interactions are not indicated. (C) Expression patterns of the six gap genes and their linearly ordered PP representation. For each of the six gap genes, the regions diagrammed in blue are the PP with sPP coefficient greater than or equal to 0.1 for at least one of the replicate images, while the regions diagrammed in white are the PP with a coefficient less than 0.1 for all replicate images. To evaluate the prediction performance of the SLCN in (A), we first mapped each node in (B) to a connected PP group in (C). According to (C), *gt* has two major components: the anterior part which has expression in PP6 and PP7, and the posterior part that has expression in PP9. The anterior gt_1 and the posterior gt_2 symbols in (B) can be mapped to these two components respectively. *hb* also has two major connected PP components: the anterior part which has expression in PP6-8 that corresponds to hb_1 in (B), and the posterior part that has expression in PP17 and PP20 that corresponds to hb_2 in (B). For *hkb*, the only expression in PP20 corresponds to the *hkb* gene symbol in (B). *kni* has two components. The first one in PP6 does not correspond to any node in (B) (the $*$ symbol indicates a region of gene expression with no match in (B)), whereas the second one in PP8 and PP9 corresponds to the *kni* symbol in (B). Similarly, the first component of *Kr* in PP8 corresponds to the symbol *Kr* in (B), whereas the posterior part in PP20 does not appear in (B). Finally, the only component of *tll* in PP17 and PP20 correspond to the only *tll* symbol in (B).

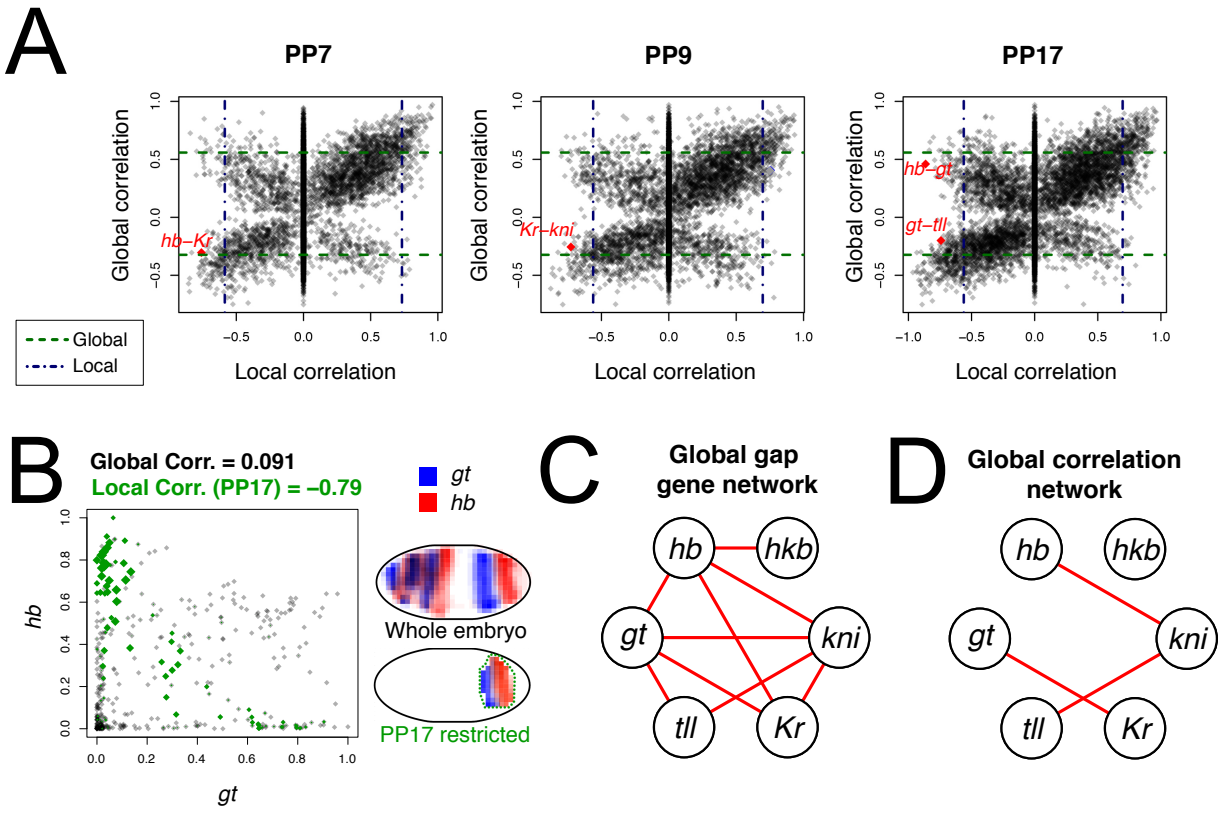

Figure 6.3: Correlating transcription factors (TF) over the whole embryo (global correlation). (**A**) Scatter plots of the global correlation vs. the local correlations for PP7, PP9 and PP17. The dashed lines correspond to the lower and upper cutoffs for the local correlations (vertical lines) and global correlations (horizontal lines). The four branches of each plot are due to the fact that the distributions of both global correlations and local correlations are bimodal. Highlighted in the scatter plots are the gap-gene links correctly identified by the local networks but missed by the global network. (**B**) The PP-based correlation approach detected locally complementary patterns whereas the global correlation approach failed. The scatterplot showed the pixel-wise intensity relationship between a pair of expression images of *gt* and *hb*. The green dots corresponded to the pixels in the region defined by PP17, with dot size proportional to the pixel intensity of PP17. We observed a clear negative association between the two TF in PP17. However, this association disappears when we consider the scatterplot of all 405 pixels of the embryo. (**C**) The gap gene network without the spatial information. Here, two gap genes share an link if they share an link in the schematic gap gene network (Figure 6.2*B*) regardless of the location of the interaction. (**D**) The gap gene constructed based on correlation measurements over the whole embryo identified only three out of nine links of the global version of gap gene network in (C).

Part II

Theoretical dictionary learning: local identifiability

Chapter 7

Theoretical dictionary learning: introduction

In the first part of this thesis, we combined a dictionary learning algorithm, nonnegative matrix factorization (NMF), with a novel stability-based criterion for model selection (staNMF), to analyze *Drosophila* embryonic spatial gene expression patterns and build local networks for transcription factors. The biological interpretability of the NMF-derived dictionary and the success of the downstream analyses motivated us to investigate why dictionary learning works. In the second part of this thesis, we will study a particular formulation of dictionary learning with the l_1-norm objective function. By considering a property called *local identifiability*, we will study when dictionary learning is a mathematically well-posed problem. A sufficient and almost necessary condition for local identifiability will be provided for two reasonable signal generation models. In this chapter, we will first give a review of dictionary learning theory and introduce a mathematical formulation of our problem.

7.1 Introduction

Expressing signals as sparse linear combinations of a dictionary basis has enjoyed great success in applications ranging from image denoising to audio compression. Given a known dictionary matrix $\mathbf{D} \in \mathbb{R}^{d\times K}$ with K columns or atoms, one popular method to recover sparse coefficients $\boldsymbol{\alpha} \in \mathbb{R}^K$ of the signal $\mathbf{x} \in \mathbb{R}^d$ is through solving the convex l_1-minimization problem:

$$\text{minimize } \|\boldsymbol{\alpha}\|_1 \text{ subject to } \mathbf{x} = \mathbf{D}\boldsymbol{\alpha}.$$

This approach, known as *basis pursuit* [66], along with many of its variants, has been studied extensively in statistics and signal processing communities. See, e.g. [67, 68, 69].

For certain data types such as natural image patches, predefined dictionaries like the wavelets [70] are usually available. However, when a less-known data type is encountered, a new dictionary has to be designed for effective representations. Dictionary learning, or sparse coding, learns adaptively a dictionary from a set of training signals such that they

have sparse representations under this dictionary [38]. One formulation of dictionary learning involves solving a non-convex l_1-minimization problem [71, 52, 53]. Concretely, define

$$l(\mathbf{x}, \mathbf{D}) = \min_{\boldsymbol{\alpha} \in \mathbb{R}^K} \{\|\boldsymbol{\alpha}\|_1, \text{ subject to } \mathbf{x} = \mathbf{D}\boldsymbol{\alpha}\}. \tag{7.1}$$

We learn a dictionary from the N signals $\mathbf{x}_i \in \mathbb{R}^d$ for $i = 1, ..., N$ by solving:

$$\min_{\mathbf{D} \in \mathcal{D}} \ L_N(\mathbf{D}) = \min_{\mathbf{D} \in \mathcal{D}} \frac{1}{N} \sum_{i=1}^{N} l(\mathbf{x}_i, \mathbf{D}). \tag{7.2}$$

Here, $\mathcal{D} \subset \mathbb{R}^{d \times K}$ is a constraint set for candidate dictionaries. In many signal processing tasks, learning an adaptive dictionary via the optimization problem (7.2) and its variants is empirically demonstrated to have superior performance over fixed standard dictionaries [72, 73, 74]. For a review of dictionary learning algorithms and applications, see [75, 76, 77].

Despite the empirical success of many dictionary learning formulations, relatively little theory is available to explain why they work. One line of research addresses the problem of *dictionary identifiability*: if the signals are generated using a dictionary $\mathbf{D}_0$ referred to as the *reference dictionary*, under what conditions can we recover $\mathbf{D}_0$ by solving the dictionary learning problem? Being able to identify the reference dictionary is important when we interpret the learned dictionary as for our spatial gene expression data. Let $\boldsymbol{\alpha}_i \in \mathbb{R}^K$ for $i = 1, ..., N$ be some random vectors. A popular signal generation model assumes that a signal vector can be expressed as a linear combination of the columns of the reference dictionary: $\mathbf{x}_i \approx \mathbf{D}_0\boldsymbol{\alpha}_i$ [52, 53, 54]. In this thesis, we will study the problem of *local identifiability* of l_1-minimization dictionary learning (7.2) under this generating model.

Local identifiability. A reference dictionary $\mathbf{D}_0$ is said to be *locally identifiable* with respect to an objective function $L(\mathbf{D})$ if $\mathbf{D}_0$ is one of the local minima of L. The pioneer work of [52] (referred to as GS henceforth) analyzed the l_1-minimization problem (7.2) for noiseless signals ($\mathbf{x}_i = \mathbf{D}_0\boldsymbol{\alpha}_i$) and complete ($d = K$ and full rank) dictionaries. Under a sparse Bernoulli-Gaussian model for the linear coefficients $\boldsymbol{\alpha}_i$'s, they showed that for a sufficiently incoherent reference dictionary $\mathbf{D}_0$, $N = O(K \log K)$ samples can guarantee local identifiability with respect to $L_N(\mathbf{D})$ in (7.2) with high probability. Still in the noiseless setting, [53] extended the analysis to over-complete ($d > K$) dictionaries. More recently under the noisy linear generative model ($\mathbf{x}_i = \mathbf{D}_0\boldsymbol{\alpha}_i + \text{noise}$) and over-complete dictionary setting, [54] developed the theory of local identifiability for (7.2) with $l(\mathbf{x}, \mathbf{D})$ replaced by the LASSO objective function of [48]. Other related works on local identifiability include [78] and [79], who gave respectively sufficient conditions for the local correctness of the K-SVD [80] algorithm and a maximum response formulation of dictionary learning.

Contributions. There has not been much work on necessary conditions for local dictionary identifiability. Numerical experiments demonstrate that there seems to be a phase boundary for local identifiability (Figure 7.1). The bound implied by the sufficient condition in GS falls well below the simulated phase boundary, suggesting that their result can be further

improved. Thus, even though theoretical results for the more general scenarios are available, we adapt the noiseless signals and complete dictionary setting of GS in order to find better local identifiability conditions. We summarize our major contributions below:

- For the population case where $N = \infty$, we establish a sufficient and almost necessary condition for local identifiability under both the s-sparse Gaussian model and the Bernoulli-Gaussian model. For the Bernoulli-Gaussian model, the phase boundary implied by our condition significantly improves the GS bound and agrees well with the simulated phase boundary (Figure 7.1).

- We provide lower and upper bounds to approximate the quantities involved in our sufficient and almost necessary condition, as it generally requires to solve a series of second-order cone programs to compute those quantities.

- As a consequence, we show that a μ-coherent reference dictionary – a dictionary with absolute pairwise column inner-product at most $\mu \in [0, 1)$ – is locally identifiable for sparsity level, measured by the average number of nonzeros in the random linear coefficient vectors, up to the order $O(\mu^{-2})$. Moreover, if the sparsity level is greater than $O(\mu^{-2})$, the reference dictionary is generally not locally identifiable. In comparison, instead of imposing condition on the sparsity level, the sufficient condition by GS demands the number of dictionary atoms $K = O(\mu^{-2})$, which is a much more stringent requirement. For over-complete dictionaries, [53] requires the sparsity level to be of the order $O(\mu^{-1})$. It should also be noted that [79] established the bound $O(\mu^{-2})$ for *approximate* local identifiability under a new response maximization formulation of dictionary learning. Our result is the first to show that $O(\mu^{-2})$ is achievable and optimal for *exact* local recovery under the l_1-minimization criterion.

- We also extend our identifiability results to the finite sample case. We show that for a fixed sparsity level, we need $N = O(K \log K)$ $i.i.d$ signals to determine whether or not the reference dictionary can be identified locally. This sample requirement is the same as GS's and is the best known sample requirement among all previous studies on local identifiability.

Other related works. Apart from analyzing the local minima of dictionary learning, another line of research aims at designing provable algorithms for recovering the reference dictionary. [81] and [82] proposed combinatorial algorithms and gave deterministic conditions for dictionary recovery which require sample size N to be exponentially large in the number of dictionary atoms K. [83] established exact global recovery results for complete dictionaries through efficient convex programs. [84] and [85] proposed clustering-based methods to estimate the reference dictionary in the overcomplete setting. [86] and [87] provided theoretical guarantees for their alternating minimization algorithms. [88] proposed a non-convex optimization algorithm that provably recovers a complete reference dictionary for sparsity level up to $O(K)$. While in this thesis we do not provide an algorithm, our identifiability

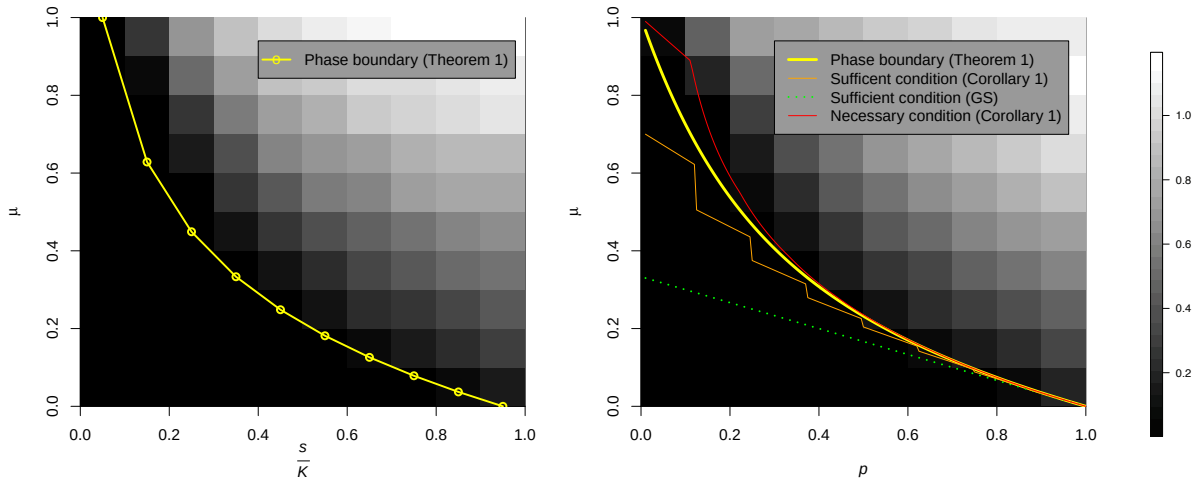

Figure 7.1: Local recovery error for the s-sparse Gaussian model (Left) and the Bernoulli(p)-Gaussian model (Right). The parameter $s \in \{1, ..., K\}$ is the number of nonzeros in each linear coefficient vector under the s-sparse Gaussian model, and $p \in (0, 1]$ is the probability of an entry of the linear coefficient vector being nonzero under the Bernoulli(p)-Gaussian model. The data are generated with the reference dictionary $\mathbf{D}_0 \in \mathbb{R}^{10\times 10}$ (i.e. $K = 10$) satisfying $\mathbf{D}_0^T\mathbf{D}_0 = \mu\mathbf{1}\mathbf{1}^T + (1-\mu)\mathbf{I}$ for $\mu \in [0, 1)$, see Example 8.5 for details. For each $(\mu, \frac{s}{K})$ or (μ, p) tuple, ten batches of $N = 2000$ signals $\{\mathbf{x}_i\}_{i=1}^{2000}$ are generated according to the noiseless linear model $\mathbf{x}_i = \mathbf{D}_0\boldsymbol{\alpha}_i$, with $\{\boldsymbol{\alpha}_i\}_{i=1}^{2000}$ drawn $i.i.d$ from the s-sparse Gaussian model or $i.i.d$ from the Bernoulli(p)-Gaussian model. For each batch, the dictionary is estimated through an alternating minimization algorithm in the SPAMS package [47], with initial dictionary set to be $\mathbf{D}_0$. The grayscale intensity in the figure corresponds to the Frobenius error of the difference between the estimated dictionary and the reference dictionary $\mathbf{D}_0$, averaged for the ten batches. The "phase boundary" curve corresponds to the theoretical boundary that separates the region of local identifiability (below the curve) and the region of local non-identifiability (above the curve) according to Theorem 1 of this thesis. The "Sufficient condition (Corollary 1)" and "Necessary condition (Corollary 1)" curves are the lower and upper bounds given by Corollary 1 to approximate the exact phase boundary. Finally, the "Sufficient condition (GS)" curve corresponds to the lower bound by GS. Note that for the s-sparse Gaussian model, the "Sufficient condition (Corollary 1)" and "Necessary condition (Corollary 1)" curves coincide with the phase boundary.

conditions suggest theoretical limits of dictionary recovery for all algorithms attempting to solve the optimization problem (7.2). In particular, in the regime where the reference dictionary is not identifiable, no algorithm can simultaneously solve (7.2) and return the ground truth reference dictionary.

Other related works include generalization bounds for signal reconstruction errors under the learned dictionary [89, 90, 91, 92], dictionary identifiability through combinatorial matrix theory [93], as well as algorithms and theories for the closely related independent component analysis [94, 95] and nonnegative matrix factorization [96, 97].

7.2 Preliminaries

Notations

For a positive integer m, define $[\![m]\!]$ to be the set of the first m positive integers, $\{1, ..., m\}$. The notation $\mathbf{x}[i]$ denotes the i-th entry of the vector $\mathbf{x} \in \mathbb{R}^m$. For a non-empty index set $S \subset [\![m]\!]$, we denote by $|S|$ the set cardinality and $\mathbf{x}[S] \in \mathbb{R}^{|S|}$ the sub-vector indexed by S. We define $\mathbf{x}[-j] \coloneqq (\mathbf{x}[1], ..., \mathbf{x}[j-1], \mathbf{x}[j+1], ..., \mathbf{x}[m]) \in \mathbb{R}^{m-1}$ to be the vector $\mathbf{x}$ without its j-th entry.

For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we denote by $\mathbf{A}[i, j]$ its (i, j)-th entry. For non-empty sets $S \subset [\![m]\!]$ and $T \subset [\![n]\!]$, denote by $\mathbf{A}[S, T]$ the submatrix of $\mathbf{A}$ with the rows indexed by S and columns indexed by T. Denote by $\mathbf{A}[i,]$ and $\mathbf{A}[, j]$ the i-th row and the j-th column of $\mathbf{A}$ respectively. Similar to the vector case, the notation $\mathbf{A}[-i, j] \in \mathbb{R}^{(m-1) \times n}$ denotes the j-th column of $\mathbf{A}$ without its i-th entry.

For $p \geq 1$, the l_p-norm of a vector $\mathbf{x} \in \mathbb{R}^m$ is defined as $\|\mathbf{x}\|_p = (\sum_{i=1}^m |\mathbf{x}[i]|^p)^{1/p}$, with the convention that $\|\mathbf{x}\|_0 = |\{i : \mathbf{x}[i] \neq 0\}|$ and $\|\mathbf{x}\|_\infty = \max_i |\mathbf{x}[i]|$. For any norm $\|.\|$ on $\mathbb{R}^m$, the dual norm of $\|.\|$ is defined as $\|\mathbf{x}\|^* = \sup_{\mathbf{y} \neq 0} \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{y}\|}$.

For two sequences of real numbers $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we denote by $a_n = O(b_n)$ if there is a constant $C > 0$ such that $a_n \leq C b_n$ for all $n \geq 1$. For $a \in \mathbb{R}$, denote by $\lfloor a \rfloor$ the integer part of a and $\lceil a \rceil$ the smallest integer greater than or equal to a. Throughout this thesis, we shall agree that $\frac{0}{0} = 0$.

Basic assumptions

We denote by $\mathcal{D} \subset \mathbb{R}^{d \times K}$ the constraint set of dictionaries for the optimization problem (7.2). In this thesis, since we focus on complete dictionaries, we assume $d = K$. As in GS, we choose $\mathcal{D}$ to be the *oblique manifold* [98]:

$$\mathcal{D} = \left\{ \mathbf{D} \in \mathbb{R}^{K \times K} : \ \|\mathbf{D}[, k]\|_2 = 1 \text{ for all } k = 1, ..., K \right\}.$$

We also call a column of the dictionary $\mathbf{D}[, k]$ an *atom* of the dictionary. Denote by $\mathbf{D}_0 \in \mathcal{D}$ the *reference dictionary* – the ground truth dictionary that generates the signals. With these notations, we now give a formal definition for local identifiability:

Definition 7.1. (Local identifiability) Let $L(\mathbf{D}) : \mathcal{D} \to \mathbb{R}$ be an objective function. We say that the reference dictionary $\mathbf{D}_0$ is *locally identifiable* with respect to $L(\mathbf{D})$ if $\mathbf{D}_0$ is a local minimum of $L(\mathbf{D})$.

Sign-permutation ambiguity. As noted by previous works GS and [53], there is an intrinsic sign-permutation ambiguity with the l_1-norm objective function $L(\mathbf{D}) = L_N(\mathbf{D})$ of (7.2). Let $\mathbf{D}' = \mathbf{DP\Lambda}$ for some permutation matrix $\mathbf{P}$ and diagonal matrix $\mathbf{\Lambda}$ with ± 1 diagonal entries. It is easy to see that $\mathbf{D}'$ and $\mathbf{D}$ have the same objective value. Thus, the objective function $L_N(\mathbf{D})$ has at least $2^n n!$ local minima. We can only recover $\mathbf{D}_0$ up to column permutation and column sign changes.

Note that if the dictionary atoms are linearly dependent, the effective dimension is strictly less than K and the problem essentially becomes over-complete. Since dealing with over-complete dictionaries is beyond the scope of this thesis, we make the following assumption:

Assumption I *(Complete dictionaries). The reference dictionary $\mathbf{D}_0 \in \mathcal{D} \subset \mathbb{R}^{K \times K}$ is full rank.*

Let $\mathbf{M}_0 = \mathbf{D}_0^T \mathbf{D}_0$ be the *dictionary atom collinearity matrix* containing the inner-products between dictionary atoms. Since each dictionary atom has unit l_2-norm, $\mathbf{M}_0[i,i] = 1$ for all $i \in [\![K]\!]$. In addition, as $\mathbf{D}_0$ is full rank, $\mathbf{M}_0$ is positive definite and $|\mathbf{M}_0[i,j]| < 1$ for all $i \neq j$.

We assume that a signal is generated as a random linear combination of the dictionary atoms. In this thesis, we consider the following two probabilistic models for the random linear coefficients:

Probabilistic models for sparse coefficients. Denote by $\mathbf{z} \in \mathbb{R}^m$ a random vector from the K-dimensional standard normal distribution.

Model 1 – *$SG(s)$.* Let $\mathbf{S}$ be a size-s subset uniformly drawn from all size-s subsets of $[\![K]\!]$. Define $\boldsymbol{\xi} \in \{0,1\}^K$ by setting $\boldsymbol{\xi}[j] = I\{j \in \mathbf{S}\}$ for $j \in [\![K]\!]$, where $I\{.\}$ is the indicator function. Let $\boldsymbol{\alpha} \in \mathbb{R}^m$ be such that $\boldsymbol{\alpha}[j] = \boldsymbol{\xi}[j]\mathbf{z}[j]$. Then we say $\boldsymbol{\alpha}$ is drawn from the *s-sparse Gaussian model*, or $SG(s)$.

Model 2 – *$BG(p)$.* For $j \in [\![K]\!]$, let $\boldsymbol{\xi}[j]$'s be *i.i.d.* Bernoulli random variable with success probability $p \in (0,1]$. Let $\boldsymbol{\alpha} \in \mathbb{R}^m$ be such that $\boldsymbol{\alpha}[j] = \boldsymbol{\xi}[j]\mathbf{z}[j]$. Then we say $\boldsymbol{\alpha}$ is drawn from the *Bernoulli(p)-Gaussian model*, or $BG(p)$.

With the above two models we can formally state the following assumption for random signal generation:

Assumption II *(Signal generation). For $i \in [\![N]\!]$, let $\boldsymbol{\alpha}_i$'s be either i.i.d. s-sparse Gaussian vectors or i.i.d. Bernoulli(p)-Gaussian vectors. The signals $\mathbf{x}_i$'s are generated according to the noiseless linear model:*

$$\mathbf{x}_i = \mathbf{D}_0 \boldsymbol{\alpha}_i.$$

Remarks:
(1) The above two models and their variants were studied in a number of prior theoretical works, including [52, 53, 54, 99, 88].
(2) By construction, a random vector generated from the s-sparse model has exactly s nonzero entries. The data points $\mathbf{x}_i$'s therefore lie within the union of the linear spans of s

dictionary atoms (Figure 7.2 Left). The Bernoulli(p)-Gaussian model, on the other hand, allows the random coefficient vector to have any number of nonzero entries ranging from 0 to K with a mean pK. As a result, the data points can be outside of any sparse linear span of the dictionary atoms (Figure 7.2 Right). We refer readers to the remarks following Example 8.5 for a discussion of the effect of non-sparse outliers on local identifiability.
(3) Our local identifiability results can be extended to a wider class of sub-Gaussian distributions. However, such an extension will lead to an increase in complexity of the form of the quantities involved in our theorems. For proof of concept, we will only focus on the standard Gaussian distribution.

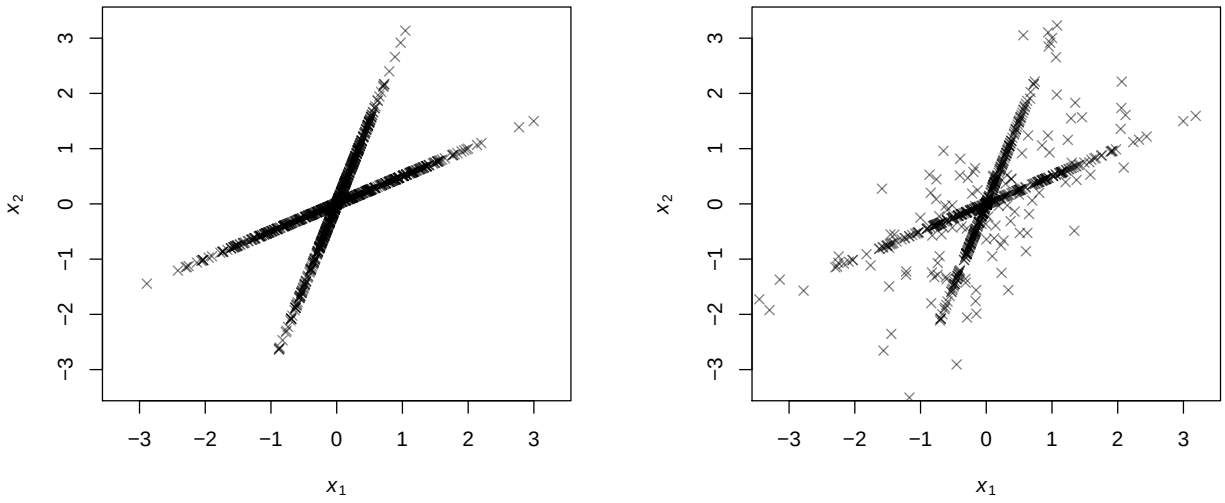

Figure 7.2: Data generation for $K = 2$. Left: the s-sparse Gaussian model with $s = 1$; Right: the Bernoulli(p)-Gaussian model with $p = 0.2$. The dictionary is constructed such that the inner product between the two dictionary atoms is 0.7. A sample of $N = 1000$ data points are generated for both models. For the s-sparse model all data points are perfectly aligned with the two lines corresponding to the two dictionary atoms. For the Bernoulli(p)-Gaussian model, a number of data points fall outside the two lines. According to our Theorems 1 and 3, despite those outliers and the high collinearity between the two atoms, the reference dictionary is still locally identifiable at the population level and with high probability for finite samples.

In this thesis, we study the problem of dictionary identifiability with respect to the population objective function $\mathbb{E}\, L_N(\mathbf{D})$ (Chapter 8) and the finite sample objective function $L_N(\mathbf{D})$ (Chapter 9). In order to analyze these objective functions, it is convenient to define the following "group LASSO"-type norms:

Definition 7.2. For an integer $m \geq 2$ and $\mathbf{w} \in \mathbb{R}^m$.

1. For $k \in [\![m]\!]$, define
$$|||\mathbf{w}|||_k = \frac{\sum_{|S|=k} \|\mathbf{w}[S]\|_2}{\binom{m-1}{k-1}}.$$

2. For $p \in (0,1)$, define
$$|||\mathbf{w}|||_p = \sum_{k=0}^{m-1} \texttt{pbinom}(k; m-1, p) |||\mathbf{w}|||_{k+1},$$
where pbinom is the probability mass function of the binomial distribution:
$$\texttt{pbinom}(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Remark:
(1) Note that the above norms $|||\mathbf{w}|||_k$ and $|||\mathbf{w}|||_p$ are in fact the expected values of $|\mathbf{w}^T\boldsymbol{\alpha}|$ with the random vector $\boldsymbol{\alpha}$ drawn from the $SG(s)$ model and the $BG(p)$ model respectively. For invertible $\mathbf{D} \in \mathcal{D}$, it can be shown that the objective function for one signal $\mathbf{x} = \mathbf{D}_0\boldsymbol{\alpha}$ is
$$l(\mathbf{x}, \mathbf{D}) = \|\mathbf{H}\boldsymbol{\alpha}\|_1 = \sum_{j=1}^{K} |\mathbf{H}[j,]\boldsymbol{\alpha}|,$$
where $\mathbf{H} = \mathbf{D}^{-1}\mathbf{D}_0$. Thus, taking the expectation of the objective function with respect to $\mathbf{x}$, we end up with a quantity involving either $\sum_{j=1}^{K} |||\mathbf{H}[j,]|||_s$ or $\sum_{j=1}^{K} |||\mathbf{H}[j,]|||_p$. This is the motivation of defining these norms.
(2) In particular, $|||\mathbf{w}|||_1 = \|\mathbf{w}\|_1$ and $|||\mathbf{w}|||_m = \|\mathbf{w}\|_2$.
(3) The norms defined above are special cases of the group LASSO penalty by [100]. For $|||\mathbf{w}|||_k$, the summation covers all size-k subsets of $[\![m]\!]$. The normalization factor is the number of times $\mathbf{w}[i]$ appears in the numerator. Thus, $|||\mathbf{w}|||_k$ is essentially the average of the l_2-norms of all size-k sub-vectors of $\mathbf{w}$. On the other hand, $|||\mathbf{w}|||_p$ is a weighted average of $|||\mathbf{w}|||_k$'s with binomial probabilities.

Chapter 8

Population local identifiability

In this chapter, we will establish local identifiability results when infinitely many signals are observed.

Denote by $\mathbb{E}\ l(\mathbf{x}_1, \mathbf{D})$ the expectation of the objective function $l(\mathbf{x}_1, \mathbf{D})$ of (7.1) with respect to the random signal $\mathbf{x}_1$. By the central limit theorem, as the number of signals N tends to infinity, the empirical objective function $L_N(\mathbf{D}) = \frac{1}{N}\sum_{i=1}^{N} l(\mathbf{x}_i, \mathbf{D})$ converges almost surely to its population mean $\mathbb{E}\ l(\mathbf{x}_1, \mathbf{D})$ for each fixed $\mathbf{D} \in \mathcal{D}$. Therefore the population version of the optimization problem (7.2) is:

$$\min_{\mathbf{D}\in\mathcal{D}} \mathbb{E}\ l(\mathbf{x}_1, \mathbf{D}) \tag{8.1}$$

Note that we only need to work with $\mathbf{D} \in \mathcal{D}$ that is full rank. Indeed, if the linear span of the columns of $\mathbf{D}$ $\text{span}(\mathbf{D}) \neq \mathbb{R}^K$, then $\mathbf{D}_0\boldsymbol{\alpha}_1 \notin \text{span}(\mathbf{D})$ with nonzero probability. Thus $\mathbf{D}$ is infeasible with nonzero probability and so $\mathbb{E}\ l(\mathbf{x}_1, \mathbf{D}) = +\infty$. For a full rank dictionary $\mathbf{D}$, the following lemma gives the closed-form expressions for the expected objective function $\mathbb{E}\ l(\mathbf{x}_1, \mathbf{D})$:

Lemma 8.1. *(Closed-form objective functions) Let* $\mathbf{D}$ *be a full rank dictionary in* $\mathcal{D}$ *and* $\mathbf{x}_1 = \mathbf{D}_0\boldsymbol{\alpha}_1$ *where* $\boldsymbol{\alpha}_1 \in \mathbb{R}^K$ *is a random vector. For notational convenience, let* $\mathbf{H} = \mathbf{D}^{-1}\mathbf{D}_0$.

1. *If* $\boldsymbol{\alpha}_1$ *is generated according to the* $SG(s)$ *model with* $s \in [\![K-1]\!]$,

$$L_{SG(s)}(\mathbf{D}) \coloneqq \mathbb{E}\ l(\mathbf{x}_1, \mathbf{D}) = \sqrt{\frac{2}{\pi}}\frac{s}{K}\sum_{j=1}^{K} |||\mathbf{H}[j,]|||_s. \tag{8.2}$$

2. *If* $\boldsymbol{\alpha}_1$ *is generated according to the* $BG(p)$ *model with* $p \in (0,1)$,

$$L_{BG(p)}(\mathbf{D}) \coloneqq \mathbb{E}\ l(\mathbf{x}_1, \mathbf{D}) = \sqrt{\frac{2}{\pi}}p\sum_{j=1}^{K} |||\mathbf{H}[j,]|||_p. \tag{8.3}$$

For the non-sparse cases $s = K$ and $p = 1$, we have

$$L_{SG(s)}(\mathbf{D}) = L_{BG(p)}(\mathbf{D}) = \sqrt{\frac{2}{\pi}} \sum_{j=1}^{K} \|\mathbf{H}[j,]\|_2.$$

Remark: It can be seen from the above closed-form expressions that the two models are closely related. First of all, it is natural to identify p with $\frac{s}{K}$, the fraction of expected number of nonzero entries in $\boldsymbol{\alpha}_1$. Next, by definition, $|||.|||_p$ is a binomial average of $|||.|||_k$. Therefore, the Bernoulli-Gaussian objective function $L_{BG(p)}(\mathbf{D})$ can be treated as a binomial average of the s-sparse objective function $L_{SG(s)}(\mathbf{D})$.

8.1 A sufficient and almost necessary condition

By analyzing the above closed-form expressions of the l_1-norm objective function, we establish the following sufficient and almost necessary conditions for population local identifiability:

Theorem 1. *(Population local identifiability) Recall that $\mathbf{M}_0 = \mathbf{D}_0^T \mathbf{D}_0$ and $\mathbf{M}_0[-j, j]$ denotes the j-th column of $\mathbf{M}_0$ without its j-th entry. Let $|||.|||_s^*$ and $|||.|||_p^*$ be the dual norm of $|||.|||_s$ and $|||.|||_p$ respectively.*

1. *(SG(s) models) For $K \geq 2$ and $s \in [\![K-1]\!]$, if*

$$\max_{j \in [\![K]\!]} |||\mathbf{M}_0[-j, j]|||_s^* < 1 - \frac{s-1}{K-1}.$$

   then $\mathbf{D}_0$ is locally identifiable with respect to $L_{SG(s)}$.

2. *(BG(p) models) For $K \geq 2$ and $p \in (0, 1)$, if*

$$\max_{j \in [\![K]\!]} |||\mathbf{M}_0[-j, j]|||_p^* < 1 - p.$$

   then $\mathbf{D}_0$ is locally identifiable with respect to $L_{BG(p)}$.

Moreover, the above conditions are almost necessary in the sense that if the reverse strict inequalities hold, then $\mathbf{D}_0$ is not locally identifiable.

On the other hand, if $s = K$ or $p = 1$, then $\mathbf{D}_0$ is not locally identifiable with respect to $L_{SG(s)}$ or $L_{BG(p)}$.

Proof sketch. Let $\{\mathbf{D}_t\}_{t \in \mathbb{R}}$ be a collection of dictionaries $\mathbf{D}_t \in \mathcal{D}$ indexed by $t \in \mathbb{R}$ and $L(\mathbf{D}) = \mathbb{E}\, l(\mathbf{x}_1, \mathbf{D})$ be the population objective function. The reference dictionary $\mathbf{D}_0$ is a local minimum of $L(\mathbf{D})$ on the manifold $\mathcal{D}$ if and only if the following statement holds: for

any $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$ that is a smooth function of t with non-vanishing derivative at $t = 0$, $L(\mathbf{D}_t)$ has a local minimum at $t = 0$. For a fixed $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$, to ensure that $L(\mathbf{D}_t)$ achieves a local minimum at $t = 0$, it suffices to have the following one-sided derivative inequalities:

$$\lim_{t\downarrow 0^+} \frac{L(\mathbf{D}_t) - L(\mathbf{D}_0)}{t} > 0 \text{ and } \lim_{t\uparrow 0^-} \frac{L(\mathbf{D}_t) - L(\mathbf{D}_0)}{t} < 0.$$

With some algebra, the two inequalities can be translated into the following statement:

$$\max_{j\in[\![K]\!]} \left|\mathbf{M}_0[-j, j]^T\mathbf{w}\right| < \begin{cases} 1 - \frac{s-1}{K-1} & \text{for } SG(s) \text{ models} \\ 1 - p & \text{for } BG(p) \text{ models} \end{cases}$$

where $\mathbf{w} \in \mathbb{R}^{K-1}$ is a unit vector in terms of the norm $|||.|||_s$ or $|||.|||_p$ and it corresponds to the "approaching direction" of $\mathbf{D}_t$ to $\mathbf{D}_0$ on $\mathcal{D}$ as t tends to zero. Since $t = 0$ has to be a local minimum for all smooth $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$ or approaching directions, by taking the supremum over all such unit vectors the LHS of the above inequality becomes the dual norm of $|||.|||_s$ or $|||.|||_p$. On the other hand, $\mathbf{D}_0$ is not a local minimum if $\lim_{t\downarrow 0^+}(L(\mathbf{D}_t) - L(\mathbf{D}_0))/t < 0$ or $\lim_{t\uparrow 0^-}(L(\mathbf{D}_t) - L(\mathbf{D}_0))/t > 0$ for some $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$. Thus our condition is also almost necessary. We refer readers to Section A.1 for the detailed proof.

Local identifiability phase boundary. The conditions in Theorem 1 indicate that population local identifiability undergoes a phase transition. The following equations

$$\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j, j]|||_s^* = 1 - \frac{s-1}{K-1} \text{ and } \max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j, j]|||_p^* = 1 - p$$

define the local identifiability phase boundaries which separate the region of local identifiability, in terms of dictionary atom collinearity matrix $\mathbf{M}_0$ and the sparsity level s or p, and the region of local non-identifiability, under respective models.

The roles of dictionary atom collinearity and sparsity. Both the dictionary atom collinearity matrix $\mathbf{M}_0$ and the sparsity parameter s or p play roles in determining local identifiability. Loosely speaking, for $\mathbf{D}_0$ to be locally identifiable, neither can the atoms of $\mathbf{D}_0$ be too linearly dependent, nor can the random coefficient vectors that generate the data be too dense. For the s-sparse Gaussian model, the quantity $\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j, j]|||_s^*$ measures the size of the off-diagonal entries of $\mathbf{M}_0$ and hence the collinearity of the dictionary atoms. In addition, that quantity also depends on the sparsity parameter s. By Lemma A.3 in the Appendix, $\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j, j]|||_s^*$ is strictly increasing with respect to s for $\mathbf{M}_0$ whose upper-triangle portion contains at least two nonzero entries (if the upper-triangle portion contains at most one nonzero entry, then the quantity does not depend on s, see Example 8.4). Similar conclusion holds for the Bernoulli-Gaussian model. Therefore, the sparser the linear coefficients, the less restrictive the requirement on dictionary atom collinearity.

On the other hand, for a fixed $\mathbf{M}_0$, by the monotonicity of $\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j, j]|||_s^*$ with respect to s, the collection of s that leads to local identifiability is of the form $s < s^*(\mathbf{M}_0)$ for some function s^* of $\mathbf{M}_0$. Similarly for the Bernoulli-Gaussian model, $p < p^*(\mathbf{M}_0)$ for some function p^* of $\mathbf{M}_0$.

8.2 Examples

Next, we will study some examples to gain more intuition for the local identifiability conditions.

Example 8.1. *(1-sparse Gaussian model) A full rank $\mathbf{D}_0$ is always locally identifiable at the population level under a 1-sparse Gaussian model. Indeed, by Corollary A.3 in the Appendix, $|||\mathbf{M}_0[-j,j]|||_1^* = \max_{i\neq j}|\mathbf{M}_0[i,j]| < 1$ for all $j \in [\![K]\!]$. Thus, a full rank dictionary $\mathbf{D}_0$ always satisfies the sufficient condition.*

Example 8.2. *($(K-1)$-sparse Gaussian model) For $j \in [\![K]\!]$, $\mathbf{M}_0[-j,j] \in \mathbb{R}^{K-1}$. Thus by Lemma A.3,*

$$|||\mathbf{M}_0[-j,j]|||_{K-1}^* = \|\mathbf{M}_0[-j,j]\|_2.$$

Therefore the phase boundary under the $(K-1)$-sparse model is

$$\max_{j\in[\![K]\!]} \|\mathbf{M}_0[-j,j]\|_2 = \frac{1}{K}.$$

Example 8.3. *(Orthogonal dictionaries) If $\mathbf{M}_0 = \mathbf{I}$, then*

$$\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j,j]|||_s^* = \max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j,j]|||_p^* = 0.$$

Therefore orthogonal dictionaries are always locally identifiable if $s < K$ or $p < 1$.

Example 8.4. *(Minimally dependent dictionary atoms) Let $\mu \in (-1,1)$. Consider a dictionary atom collinearity matrix $\mathbf{M}_0$ such that $\mathbf{M}_0[1,2] = \mathbf{M}_0[2,1] = \mu$ and $\mathbf{M}_0[i,j] = 0$ for all other $i \neq j$. By Corollary A.4 in the Appendix,*

$$\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j,j]|||_s^* = \max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j,j]|||_p^* = |\mu|.$$

Thus the phase boundaries under respective models are:

$$|\mu| = 1 - \frac{s-1}{K-1} \text{ and } |\mu| = 1 - p.$$

Notice that when $K = 2$ and for the Bernoulli-Gaussian model, the phase boundary agrees well with the empirical phase boundary in the simulation result by GS (Figure 3 of the GS paper).

Example 8.5. *(Constant inner-product dictionaries) Let $\mathbf{M}_0 = \mu \mathbf{1}\mathbf{1}^T + (1-\mu)\mathbf{I}$, i.e. $\mathbf{D}_0[,i]^T\mathbf{D}_0[,j] = \mu$ for $1 \leq i < j \leq K$. Note that $\mathbf{M}_0$ is positive definite if and only if $\mu \in (-\frac{1}{K-1}, 1)$. By Corollary A.5 in the Appendix, we have*

$$|||\mathbf{M}_0[-j,j]|||_s^* = \sqrt{s}|\mu|.$$

Thus for the s-sparse model, the phase boundary is

$$\sqrt{s}|\mu| = 1 - \frac{s-1}{K-1}.$$

Similarly for the Bernoulli(p)-Gaussian model, we have

$$\|\|\mathbf{M}_0[-j,j]\|\|_p^* = |\mu|p(K-1)\Big(\sum_{k=0}^{K-1} \texttt{pbinom}(k, K-1, p)\sqrt{k}\Big)^{-1}.$$

Thus the phase boundary is

$$|\mu| = \frac{1-p}{p(K-1)} \sum_{k=0}^{K-1} \texttt{pbinom}(k, K-1, p)\sqrt{k}.$$

Figures 8.1 shows the phase boundaries for different dictionary sizes under the two models. As K increases, the phase boundary moves towards the lower left of the region. This observation indicates that recovering the reference dictionary locally becomes increasingly difficult for larger dictionary size.

The effect of non-sparse outliers. Example 8.5 demonstrates how the presence of non-sparse outliers in the Bernoulli-Gaussian model (Figure 7.2 Right) affects the requirements for local identifiability. Set $p = \frac{s}{K}$ in order to have the same level of sparsity with the $SG(s)$ model. Applying Jensen's inequality, one can show that

$$\frac{1-p}{p(K-1)} \sum_{k=0}^{K-1} \texttt{pbinom}(k, K-1, p)\sqrt{k} < \frac{1}{\sqrt{s}}(1 - \frac{s-1}{K-1}),$$

indicating that the phase boundary of the s-sparse models is always above that of the Bernoulli-Gaussian model with the same level of sparsity. The difference between the two phase boundaries is the extra price one has to pay, in terms of the collinearity parameter μ, for recovering the dictionary locally in the presence of non-sparse outliers. One extreme example is the case where $s = 1$ and correspondingly $p = \frac{1}{K}$. By Example 8.1, under a 1-sparse model the reference dictionary $\mathbf{D}_0$ is always locally identifiable if $|\mu| < 1$. But for the $BG(\frac{1}{K})$ model, by the remark in Corollary 1, $\mathbf{D}_0$ is not locally identifiable if $|\mu| > 1 - \frac{1}{K}$. Hence, the requirement for μ in the presence of outliers is at least $\frac{1}{K}$ more stringent than that in the case of no outliers.

However, such a difference diminishes as the number of dictionary atoms K increases. Indeed, by Lemma 8.2, one can show the following lower bound for the phase boundary of under the $BG(p)$ model:

$$\frac{1-p}{p(K-1)} \sum_{k=0}^{K-1} \texttt{pbinom}(k, K-1, p)\sqrt{k} \geq \frac{1-p}{\sqrt{p(K-1)+1}} \approx \frac{1}{\sqrt{s}}(1 - \frac{s-1}{K-1}),$$

for fixed sparsity level $p = \frac{s}{K}$ and large K.

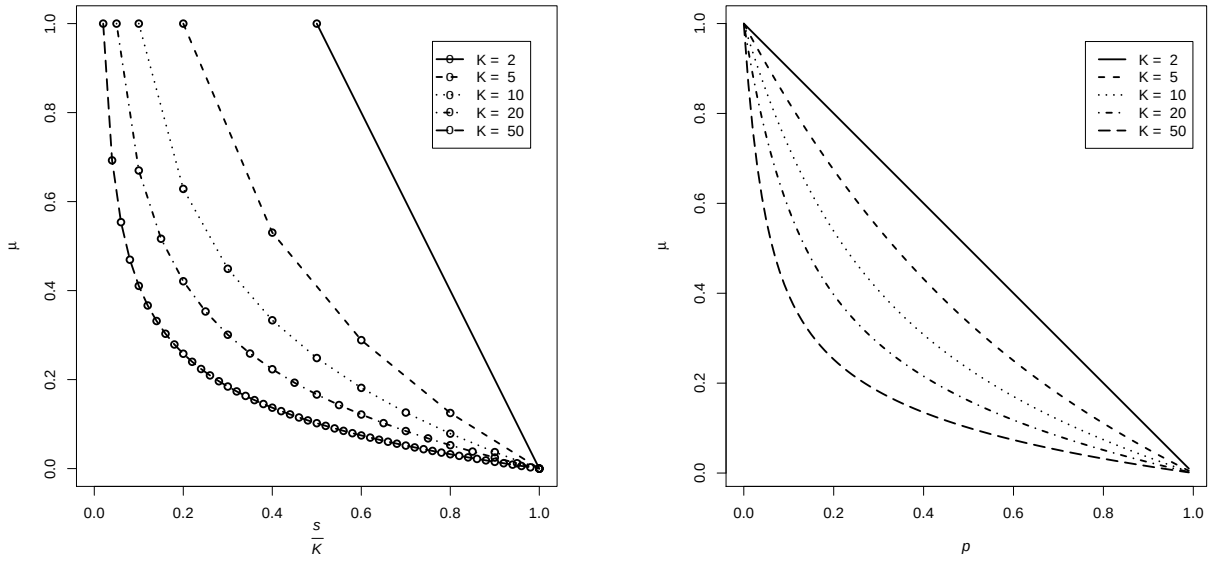

Figure 8.1: Local identifiability phase boundaries for constant inner-product dictionaries, under Left: the s-sparse Gaussian model; Right: the Bernoulli(p)-Gaussian model. For each model, phase boundaries for different dictionary sizes K are shown. Note that $\frac{s}{K} \in \{\frac{1}{K}, \frac{2}{K}, \cdots, 1\}$ and $p \in (0, 1]$. The area under the curves is the region where the reference dictionaries are locally identifiable at the population level. Due to symmetry, we only plot the portion of the phase boundaries for $\mu > 0$.

8.3 Approximation bounds

In general, the dual norms $|||.|||_s^*$ and $|||.|||_p^*$ have no closed-form expressions. According to Corollary A.2 in the Appendix, computing those quantities involves solving a second order cone problem (SOCP) with a combinatoric number of constraints. The following Lemma 8.2, on the other hand, gives computationally inexpensive approximation bounds.

Definition 8.1. (Hyper-geometric distribution related quantities) Let m be a positive integer and $d, k \in \{0\} \cup [\![m]\!]$. Denote by $L_m(d, k)$ the hypergeometric random variable with parameter m, d and k, i.e. the number of 1's after drawing without replacement k elements from d 1's and $m - d$ 0's. Now for each $d \in \{0\} \cup [\![m]\!]$, define the function $\tau_m(d, .)$ with domain on $[0, m]$ as follows: set $\tau_m(d, 0) = 0$. For $a \in (k - 1, k]$ where $k \in [\![m]\!]$, define

$$\tau_m(d, a) = \mathbb{E}\sqrt{L_m(d, k-1)} + (\mathbb{E}\sqrt{L_m(d, k)} - \mathbb{E}\sqrt{L_m(d, k-1)})(a - (k-1)).$$

Lemma 8.2. *(Lower and upper bounds for $|||.|||_s^*$ and $|||.|||_p^*$) Let m be a positive integer and $\mathbf{z} \in \mathbb{R}^m$.*

1. *For $s \in [\![m]\!]$,*

$$\max\left(\|\mathbf{z}\|_\infty, \sqrt{\frac{s}{m}} \max_{T\subset[\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\sqrt{|T|}}\right) \le \frac{s}{m} \max_{T\subset[\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\tau_m(|T|,s)} \le |||\mathbf{z}|||_s^* \le \max_{S\subset[\![m]\!],|S|=s} \|\mathbf{z}[S]\|_2.$$

2. *For $p \in (0,1)$,*

$$\max\left(\|\mathbf{z}\|_\infty, \sqrt{p} \max_{T\subset[\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\sqrt{|T|}}\right) \le p \max_{T\subset[\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\tau_m(|T|,pm)} \le |||\mathbf{z}|||_p^* \le \max_{S\subset[\![m]\!],|S|=k} \|\mathbf{z}[S]\|_2.$$

where $k = \lceil p(m-1)+1 \rceil$.

Remark:
(1) We refer readers to Lemma A.8 and A.9 for the detailed version of the above results.
(2) Since we agree that $\frac{0}{0} = 0$, the case where $T = \emptyset$ does not affect taking the maximum of all subsets.
(3) Consider a sparse vector $\mathbf{z} = (z, 0, ..., 0)^T \in \mathbb{R}^m$. By Corollary A.4,

$$|||\mathbf{z}|||_s^* = |||\mathbf{z}|||_p^* = |z| = \|\mathbf{z}\|_\infty = \max_{S\subset[\![m]\!],|S|=1} \|\mathbf{z}[S]\|_2.$$

So all the bounds are achievable by a sparse vector.
(4) Now consider a dense vector $\mathbf{z} = (z, ..., z)^T \in \mathbb{R}^m$. By Corollary A.5,

$$|||\mathbf{z}|||_s^* = \sqrt{s}|z| = \sqrt{\frac{s}{m}} \max_{T\subset[\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\sqrt{|T|}} = \max_{S\subset[\![m]\!],|S|=s} \|\mathbf{z}[S]\|_2.$$

Thus the bounds for $|||\mathbf{z}|||_s^*$ can also be achieved by a dense vector. Similarly, by the upper-bound for $|||\mathbf{z}|||_p^*$,

$$|||\mathbf{z}|||_p^* \le \sqrt{pm+1}|z|.$$

On the other hand,

$$|||\mathbf{z}|||_p^* \ge \sqrt{p} \max_{T\subset[\![m]\!]} \frac{\|\mathbf{z}\|_1}{\sqrt{|T|}} = \sqrt{p}|z| \max_{T\subset[\![m]\!]} \sqrt{|T|} = \sqrt{pm}|z|.$$

Thus both bounds for $|||\mathbf{z}|||_p^*$ are basically the same for large pm.
(5) **Computation.** To compute the lower and upper bounds efficiently, we first sort the elements of $|\mathbf{z}|$ in descending order. Without loss of generality, we can assume that $|\mathbf{z}[1]| \ge |\mathbf{z}[2]| \ge ... \ge |\mathbf{z}[m]|$. Thus the upper-bound quantity becomes

$$\max_{S\subset[\![m]\!],|S|=k} \|\mathbf{z}[S]\|_2 = (\sum_{i=1}^{k} \mathbf{z}[i]^2)^{1/2}.$$

For the lower-bound quantities, note that

$$\max_{T\subset[\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\tau_m(|T|,k)} = \max_{d\in[\![m]\!]} \max_{T\subset[\![m]\!],|T|=d} \frac{\sum_{i=1}^d |\mathbf{z}[i]|}{\tau_m(d,k)} = \max_{d\in[\![m]\!]} \frac{\sum_{i=1}^d |\mathbf{z}[i]|}{\tau_m(d,k)}.$$

Thus, the major computation burden now is to compute $\tau_m(d,k) = \mathbb{E}\sqrt{L_m(d,k)}$, for all $d \in [\![m]\!]$. We do not know a closed-form formula for $\mathbb{E}\sqrt{L_m(d,k)}$ except for $d = 1$ or $d = m$. In practice, we compute $\mathbb{E}\sqrt{L_m(d,k)}$ using its definition formula. On an OS X laptop with 1.8 GHz Intel Core i7 processor and 4GB of memory, the function `dhyper` in the statistics software `R` can compute $\mathbb{E}\sqrt{L_{2000}(d,1000)}$ for all $d \in [\![2000]\!]$ within 0.635 second. Note that the number of dictionary atoms in most applications is usually smaller than 2000.

In case m is too large, the LHS lower bounds can be used. Note that

$$\max_{T\subset[\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\sqrt{|T|}} = \max_{d\in[\![m]\!]} \frac{\sum_{i=1}^d |\mathbf{z}[i]|}{\sqrt{d}},$$

which can be computed easily.

For notational simplicity, we will define the following quantities that are involved in Lemma 8.2:

Definition 8.2. For $a \in (0, K)$, define

$$\nu_a(\mathbf{M}_0) = \max_{1\le j\le K} \max_{S\subset[\![K]\!],j\notin S} \frac{\|\mathbf{M}_0[S,j]\|_1}{\tau_{K-1}(|S|,a)}.$$

Definition 8.3. (Cumulative coherence) For $k \in [\![K-1]\!]$, define the *k-th cumulative coherence* of a reference dictionary $\mathbf{D}_0$ as

$$\mu_k(\mathbf{M}_0) = \max_{1\le j\le K} \max_{S\subset[\![K]\!],|S|=k,j\notin S} \|\mathbf{M}_0[S,j]\|_2.$$

Remark: The above quantity is actually the l_2 analog of the l_1 k-th cumulative coherence defined in [54]. Also, notice that $\mu_1(\mathbf{M}_0) = \max_{l\neq j} |\mathbf{M}_0[l,j]|$ which is the plain mutual coherence of the reference dictionary.

With the above definitions and as a direct consequence of the above Lemma 8.2, we obtain a sufficient condition and a necessary condition for population local identifiability:

Corollary 1. *Under the notations of Theorem 1, we have*

1. *Let $K \ge 2$ and $s \in [\![K-1]\!]$.*

   - *If $\mu_s(\mathbf{M}_0) < 1 - \frac{s-1}{K-1}$, then $\mathbf{D}_0$ is locally identifiable with respect to $L_{SG(s)}$;*

- *If* $\frac{s}{K-1}\nu_s(\mathbf{M}_0) > 1 - \frac{s-1}{K-1}$, *then* $\mathbf{D}_0$ *is not locally identifiable with respect to* $L_{SG(s)}$.

2. *Let* $K \geq 2$ *and* $p \in (0,1)$.

- *If* $\mu_k(\mathbf{M}_0) < 1-p$, *where* $k = \lceil p(K-2)+1 \rceil$, *then* $\mathbf{D}_0$ *is locally identifiable with respect to* $L_{BG(p)}$;
- *If* $p\nu_k(\mathbf{M}_0) > 1-p$, *where* $k = p(K-1)$, *then* $\mathbf{D}_0$ *is not locally identifiable with respect to* $L_{BG(p)}$.

Remark:

(1) In particular, by Lemma 8.2, if $\mu_1(\mathbf{M}_0) > 1 - \frac{s-1}{K-1}$ or $\mu_1(\mathbf{M}_0) > 1 - p$, then $\mathbf{D}_0$ is not locally identifiable.
(2) We can also replace $\frac{s}{K-1}\nu_s(\mathbf{M}_0)$ or $p\nu_k(\mathbf{M}_0)$ by the corresponding lower bound quantities in Lemma 8.2 which are easier to compute but give weaker necessary conditions.

Comparison with GS. Corollary 1 allows us to compare our local identifiability condition directly with that of GS. For the Bernoulli(p)-Gaussian model, the population version of the sufficient condition for local identifiability by GS is:

$$\mu_{K-1}(\mathbf{M}_0) = \max_{1 \leq j \leq K} \|\mathbf{M}_0[-j, j]\|_2 < 1 - p. \tag{8.4}$$

Note that $\mu_{K-1}(\mathbf{M}_0) \geq \mu_k(\mathbf{M}_0)$ for $k \leq K-1$.

Thus, our local identifiability result implies that of GS. Moreover, the quantity $\|\mathbf{M}_0[-j, j]\|_2$ in inequality (8.4) computes the l_2-norm of the entire $\mathbf{M}_0[-j, j]$ vector and is independent of the sparsity parameter p. On the other hand, in our sufficient condition $\max_{|S|=k, j \notin S} \|\mathbf{M}_0[S, j]\|_2$ computes the largest l_2-norm of all size-k sub-vectors of $\mathbf{M}_0[-j, j]$. Since $k = \lceil p(K-2)+1 \rceil$ is essentially pK, in the case where the model is sparse and the dictionary atom collinearity matrix $\mathbf{M}_0$ is dense, the sufficient bound by GS is most conservative compared to ours.

More concretely, let us consider constant inner-product dictionaries with parameter $\mu > 0$ as in Example 8.5. The sufficient condition by GS and the sufficient condition given by Corollary 1 are respectively

$$\sqrt{K}\mu \leq 1 - p \text{ and } \sqrt{pK+1}\mu \leq 1 - p,$$

showing that the sufficient condition by GS is much more conservative for small value of p. See Figure 7.1 for a graphical comparison of the bounds for $K = 10$.

Local identifiability for sparsity level $O(\mu^{-2})$. For notational convenience, let $\mu = \mu_1(\mathbf{M}_0)$ be the mutual coherence of the reference dictionary. For the s-sparse model, by Lemma 8.2, $\mu_s(\mathbf{M}_0) \leq \sqrt{s}\mu$. Thus the first part of the corollary implies a simpler sufficient condition:

$$\sqrt{s}\mu < 1 - \frac{s-1}{K-1}.$$

From the above inequality, it can be seen that if $1 - \frac{s-1}{K-1} > \delta$ for some $\delta > 0$, the reference dictionary is locally identifiable for sparsity level s up to the order $O(\mu^{-2})$.

Similarly for the Bernoulli(p)-Gaussian model, since

$$\mu_k(\mathbf{M}_0) \leq \sqrt{pK+1}\mu,$$

we have the following sufficient condition for local identifiability:

$$\sqrt{pK+1}\mu \leq 1 - p.$$

As before, if $1 - p > \delta$ for some $\delta > 0$, the reference dictionary is locally identifiable for sparsity level pK up to the order $O(\mu^{-2})$. On the other hand, the condition by GS requires $K = O(\mu^{-2})$, which does not take advantage of sparsity.

In addition, by Example 8.5 and the remark under Lemma 8.2, we also know that the sparsity requirement $O(\mu^{-2})$ cannot be improved in general.

Our result seems to be the first to demonstrate $O(\mu^{-2})$ is the optimal order of sparsity for exact local recovery of a reference dictionary. For a predefined over-complete dictionary, classical results such as [67] and [68] show that basis pursuit recovers an s-sparse linear coefficient vector with sparsity level s up to the order $O(\mu^{-1})$. For over-complete dictionary learning, [53] showed that exact local recovery is also possible for s-sparse model with s up to $O(\mu^{-1})$. While our results are only for complete dictionaries, we conjecture that $O(\mu^{-2})$ is also the optimal order of sparsity level for over-complete dictionaries. In fact, [79] proved that the response maximization criterion – an alternative formulation of dictionary learning – can approximately recover the over-complete reference dictionary locally with sparsity level s up to $O(\mu^{-2})$. It will be of interest to investigate whether the same sparsity requirement hold for the l_1-minimization dictionary learning (7.2) in the case of exact local recovery and over-complete dictionaries.

Chapter 9

Finite sample local identifiability

In this chapter, we will present finite sample results for local dictionary identifiability. For notational convenience, we first define the following quantities:

$$\mathcal{P}_1(\epsilon, N; \mu, K) = 2\exp\left(-\frac{N\epsilon^2}{108K\mu}\right),$$
$$\mathcal{P}_2(\epsilon, N; p, K) = 2\exp\left(-p\frac{N\epsilon^2}{18p^2K + 9\sqrt{2pK}}\right),$$
$$\mathcal{P}_3(\epsilon, N; p, K) = 3\left(\frac{24}{\epsilon p} + 1\right)^K \exp\left(-p\frac{N\epsilon^2}{360}\right).$$

Recall that $\mathbf{M}_0 = \mathbf{D}_0^T\mathbf{D}_0$ and $\mu_1(\mathbf{M}_0)$ is the mutual coherence of the reference dictionary $\mathbf{D}_0$. The following two theorems give local identifiability conditions under the s-sparse Gaussian model and the Bernoulli-Gaussian model:

Theorem 2. *(Finite sample local identifiability for $SG(s)$ models) Let $\boldsymbol{\alpha}_i \in \mathbb{R}^K$, $i \in [\![N]\!]$, be i.i.d $SG(s)$ random vectors with $s \in [\![K-1]\!]$. The signals $\mathbf{x}_i$'s are generated as $\mathbf{x}_i = \mathbf{D}_0\boldsymbol{\alpha}_i$. Assume $0 < \epsilon \leq \frac{1}{2}$,*

1. *If*

$$\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j, j]|||_s^* \leq 1 - \frac{s-1}{K-1} - \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ is locally identifiable with respect to $L_N(\mathbf{D})$ with probability exceeding

$$1 - K^2\left(\mathcal{P}_1(\epsilon, N; \mu_1(\mathbf{M}_0), K) + \mathcal{P}_2(\epsilon, N; \frac{s}{K}, K) + \mathcal{P}_3(\epsilon, N; \frac{s}{K}, K)\right).$$

2. *If*

$$\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j, j]|||_s^* \geq 1 - \frac{s-1}{K-1} + \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ *is not locally identifiable with respect to* $L_N(\mathbf{D})$ *with probability exceeding*

$$1 - K\left(\mathcal{P}_1(\epsilon, N; \mu_1(\mathbf{M}_0), K) + \mathcal{P}_2(\epsilon, N; \frac{s}{K}, K) + \mathcal{P}_3(\epsilon, N; \frac{s}{K}, K)\right).$$

Theorem 3. *(Finite sample local identifiability for BG(p) models) Let* $\boldsymbol{\alpha}_i \in \mathbb{R}^K$, $i \in [\![N]\!]$, *be i.i.d BG(p) random vectors with* $p \in (0,1)$. *The signals* $\mathbf{x}_i$*'s are generated as* $\mathbf{x}_i = \mathbf{D}_0\boldsymbol{\alpha}_i$. *Let* $K_p = K + 2p^{-1}$ *and assume* $0 < \epsilon \leq \frac{1}{2}$,

1. *If*

$$\max_{j \in [\![K]\!]} |||\mathbf{M}_0[-j, j]|||_p^* \leq 1 - p - \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ *is locally identifiable with respect to* $L_N(\mathbf{D})$ *with probability exceeding*

$$1 - K^2\left(\mathcal{P}_1(\epsilon, N; \mu_1(\mathbf{M}_0), K_p) + \mathcal{P}_2(\epsilon, N; p, K_p) + \mathcal{P}_3(\epsilon, N; p, K)\right).$$

2. *If*

$$\max_{j \in [\![K]\!]} |||\mathbf{M}_0[-j, j]|||_p^* \geq 1 - p + \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ *is not locally identifiable with respect to* $L_N(\mathbf{D})$ *with probability exceeding*

$$1 - K\left(\mathcal{P}_1(\epsilon, N; \mu_1(\mathbf{M}_0), K_p) + \mathcal{P}_2(\epsilon, N; p, K_p) + \mathcal{P}_3(\epsilon, N; p, K)\right).$$

The conditions for finite sample local identifiability are essentially identical as their population counterparts. The only difference is a margin of $\sqrt{\frac{\pi}{2}}\epsilon$ on the RHS of the inequalities. Such a margin appears in the conditions because of our proof techniques: we show that the derivative of L_N is within $O(\epsilon)$ of its expectation and then impose conditions on the expectation.

Sample size requirement. The theorems indicate that if the number of signals is a multiple of the following quantity,

$$\text{For } SG(s)\text{: } \frac{1}{\epsilon^2}\max\left\{\mu_1(\mathbf{M}_0)K\log K,\ s\log K,\ \frac{K}{s}K\log\left(\frac{K}{\epsilon s}\right)\right\}$$

$$\text{For } BG(p)\text{: } \frac{1}{\epsilon^2}\max\left\{\mu_1(\mathbf{M}_0)K\log K,\ pK\log K,\ \frac{1}{p}K\log\left(\frac{1}{\epsilon p}\right)\right\}$$

then with high probability we can determine whether or not $\mathbf{D}_0$ is locally identifiable. For ease of analysis, let us now treat ϵ as a constant. Thus, in the worst case, the sample size requirements for the two models are, respectively,

$$O(\frac{K\log K}{\frac{s}{K}}) \text{ and } O(\frac{K\log K}{p}).$$

Apart from playing a role in determining whether $\mathbf{D}_0$ is locally identifiable, the sparsity parameters s and p also affect the sample size requirement. As discussed in the population results, the sparser the linear coefficient $\boldsymbol{\alpha}_i$, the less constraint on the dictionary atom collinearity. However, with finite samples, more signals are needed to guarantee the validity of the local identifiability conditions for sparse models.

Our sample size requirement is similar to that of GS, who shows that $O(\frac{K \log K}{p(1-p)})$ signals is enough for locally recovering an incoherent reference dictionary. Our result indicates the $1-p$ factor in the denominator can be removed.

The following two corollaries are the finite sample counterparts of Corollary 1.

Corollary 2. *Under the same assumptions of Theorem 2,*

1. *(Sufficient condition for $SG(s)$ models) If*

$$\mu_s(\mathbf{M}_0) \leq 1 - \frac{s-1}{K-1} - \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ is locally identifiable with respect to $L_N(\mathbf{D})$, with the same probability bound in the first part of Theorem 2.

2. *(Necessary condition for $SG(s)$ models) If*

$$\frac{s}{K-1}\nu_s(\mathbf{M}_0) \geq 1 - \frac{s-1}{K-1} + \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ is not locally identifiable with respect to $L_N(\mathbf{D})$, with the same probability bound in the second part of Theorem 2.

Corollary 3. *Under the same assumptions of Theorem 3,*

1. *(Sufficient condition for $BG(p)$ models) Let $k = \lceil p(K-1)+1 \rceil$. If*

$$\mu_k(\mathbf{M}_0) \leq 1 - p - \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ is locally identifiable with respect to $L_N(\mathbf{D})$, with the same probability bound in the first part of Theorem 3.

2. *(Necessary condition for $BG(p)$ models) Let $k = p(K-1)$. If*

$$p\nu_k(\mathbf{M}_0) \geq 1 - p + \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ is not locally identifiable with respect to $L_N(\mathbf{D})$, with the same probability bound in the second part of Theorem 3.

Remark: As before, denote by $\mu \in [0, 1)$ the coherence of the reference dictionary. The above two corollaries together with the remark under Corollary 1 indicate that the reference dictionary is locally identifiable with high probability for sparsity level s or pK up to the order $O(\mu^{-2})$.

Proof sketch for Theorem 2 and 3. Similar to the population case, by taking one-sided derivatives of $L_N(\mathbf{D}_t)$ with respect to t at $t = 0$ for all smooth $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$, we derive a sufficient and almost necessary algebraic condition for the reference dictionary $\mathbf{D}_0$ to be a local minimum of $L_N(\mathbf{D})$. Using the concentration inequalities in Lemma A.1 - A.3, we show that the random quantities involved in the algebraic condition are close to their expectations with high probability. The population results for local identifiability can then be applied. The proofs for the two signal generation models are conceptually the same after establishing Lemma A.6 to relate the $|||.|||_p^*$ norm to the $|||.|||_s^*$ norm. The detailed proof can be found in Section A.2.

Comparison with the proof by GS. The key difference between our analysis and that of GS is that we use an alternative but equivalent formulation of dictionary learning. Instead of (7.2), GS studied the following problem:

$$\min_{\mathbf{D}\in\mathcal{D},\boldsymbol{\alpha}_i} \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{\alpha}_i\|_1 \tag{9.1}$$
$$\text{subject to } \mathbf{x}_i = \mathbf{D}\boldsymbol{\alpha}_i \text{ for all } i \in [\![N]\!].$$

Note that the above formulation optimizes jointly over $\mathbf{D}$ and $\boldsymbol{\alpha}_i$ for $i \in [\![N]\!]$, as opposed to optimizing with respect to the only parameter $\mathbf{D}$ in our case. For complete dictionaries, this formulation is equivalent to the formulation in (7.2) in the sense that $\hat{\mathbf{D}}$ is a local minimum of (7.2) if and only if $(\hat{\mathbf{D}}, \hat{\mathbf{D}}^{-1}[\mathbf{x}_1, ..., \mathbf{x}_N])$ is a local minimum of (9.1), see Remark 3.1 of GS. The number of parameters to be estimated in (9.1) is $(K-1)K + KN$, compared to $(K-1)K$ free parameters in (7.2). The growing number of parameters make the formulation employed by GS less tractable to analyze under a signal generation model.

GS did not study the population case. In their analysis, GS first obtained an algebraic condition for local identifiability that is sufficient and almost necessary. However, their condition is convoluted due to its direct dependence on the signals $\mathbf{x}_i$'s. In order to make their condition more explicit in terms of dictionary atom collinearity and sparsity level, they then investigated the condition under the Bernoulli-Gaussian model. During the probabilistic analysis, the sharp algebraic condition was weakened, resulting in a sufficient condition that is far from being necessary.

In contrast, we start with probabilistic generative models. The number of parameters is not growing as N increases, which, allows us to study the population problem directly and to apply concentration inequalities for the finite sample problem. There is little loss of information during the process of obtaining identifiability results from first principles. Therefore, studying the optimization problem (7.2) instead of (9.1) is the key to establishing an interpretable sufficient and almost necessary local identifiability condition.

Chapter 10

Conclusions

10.1 Summary

In this thesis, we proposed stability-based nonnegative matrix factorization (staNMF) to decompose spatial gene expression patterns into local principal patterns (PP). When applied to *Drosophila* embryonic expression data at early-stage, staNMF identified 21 PP that correspond to pre-organ regions, providing an informative representation of spatial gene expression patterns. We demonstrated that PP are a data-driven alternative to manual curation and facilitate the categorization of gene expression patterns. Our PP-based sparse representations (sPP) reduce large datasets to manageable scales. They allow suitable human interrogation and downstream computation on desktop computers while preserving quantitative relationships of full datasets. In addition, staNMF's utility was further substantiated by the agreement between our PP-based spatially local networks and the well-studied gap gene network.

Inspired by the success of NMF on our spatial gene expression data, we proceeded to understand why dictionary learning works. We analyzed a dictionary learning formulation with the l_1-norm objective function. In the case of noiseless signals and complete dictionaries, we established a sufficient and almost necessary condition for population local dictionary identifiability under both the s-sparse model and the Bernoulli-Gaussian model. For finite samples, we showed that as long as the number of $i.i.d$ signals scales as $O(K \log K)$, similar local identifiability conditions hold with high probability.

10.2 Future directions

Given the success of our approach for early stage *Drosophila* embryos, we expect this method to be applicable to derive meaningful data-driven representations for other data. Currently, we are extending the staNMF analysis to spatial gene expression data from later stage *Drosophila* embryos. In our preliminary analysis, staNMF was applied to about 700 segmented and aligned late stage *Drosophila* hindguts. We identified a set of PP that compart-

mentalize the developing gut and match some previously described areas of differentiation [101, 102]. Moreover, the utility of staNMF is not limited to spatial data. We used staNMF to analyze RNA-seq measurements of *Drosophila* genes for different organs, developmental stages and toxin exposures. The learned PP revealed activation of ovary and testis specific genes as a result of toxin exposure, suggesting a novel link between toxins and epigenetic inheritance. In a separate project, we collected heart specific enhancer data from mouseEncode [103], FANTOM5 [104], including DNA methylation, DNAS-seq and histone modification measurements. Using staNMF, we identified histone marks such as H3K27ac, H3K4me as key features defining active mouse heart enhancers. In addition, a random forest classifier performed nearly equally well for predicting active enhancers from raw data as from the PP representation, further demonstrating NMF as an efficient and biologically meaningful tool for dimension reduction.

Another research direction is stability analysis of the learned PP. Like all other data acquisition procedures, our image collection and preprocessing pipeline suffers from a variety of noise ranging from imprecise staging to registration artifacts. In addition, the alternating minimization implementation of NMF is intrinsically a random algorithm due to random starting values and stochastic gradient descent [47]. These two sources of randomness will ultimately affect the quality of the learned PP as well as any PP-based downstream analyses. In order to quantify the uncertainty of the learned PP, it is necessary to design a statistical inference procedure that takes into account the two sources of randomness. One possible approach is to introduce suitable data perturbations to the original data (e.g., shifts and local distortions of the expression images, adding noise) and combine them with various NMF initialization schemes. PP learned from each perturbation+NMF initialization can be matched using a greedy matching algorithm and the variability for each matched PP can be computed as our uncertain measure. In addition, this uncertainty measure can also be propagated into our PP-based gene categorization and spatially local network construction. We hope that research towards this direction can enable researchers to understand the variability of the learned data representation and subsequent analyses.

With an empirically defined cutoff, we were able to reconstruct the known *Drosophila* gap gene regulatory network using PP-based local correlations. In practice, subnetworks from a larger biological network are usually known. How do we incorporate this prior information when building networks? In recent social network research, a technique called link prediction is used to construct networks in a supervised fashion [105, 106, 107]. The idea is to treat a known interaction as positive class and a known non-interaction as negative class, and train a binary classifier such as support vector machine or random forest. The feature vector can include our local correlation and any other pairwise distance metrics between a pair of genes. By allowing an easy integration of additional information, this approach is expected to yield better results for constructing local gene networks.

There are also several directions for further development of dictionary learning theory. First of all, in this thesis we only focused on the local behaviors of the l_1-norm objective function. As pointed out by [52], numerical experiments in two dimensions suggested that local minima are in fact global minima, see Figure 2 of [52]. Thus, it is of interest to

investigate whether the conditions developed in this thesis for local identifiability are also sufficient and almost necessary for global identifiability.

Moreover, one can extend our results to a wider class of sub-Gaussian distributions other than the standard Gaussian distribution considered in this thesis. We foresee little technical difficulties for this extension. However, it should be noted that the quantities involved in our local identifiability conditions, i.e., the $|||.|||_s^*$ and $|||.|||_p^*$ norms, are consequences of the standard Gaussian assumption. Under a different distribution, it can be even more challenging to compute and approximate those quantities.

Next, it would also be desirable to improve the sufficient conditions of [53] and [54] for over-complete dictionaries and noisy data. One of the implications of our identifiability condition is that local recovery is possible for sparsity level up to order $O(\mu^{-2})$ for a μ-coherent reference dictionary. We conjecture the same sparsity requirement holds for the over-complete and/or the noisy signal case. In either case, the closed-form expression for the objective function is no longer available. A full characterization of local dictionary identifiability requires us to develop new techniques to analyze the local behaviors of the objective function.

We are also planning to establish similar identifiability results for NMF. Prior work on NMF identifiability gave sufficient conditions that are usually too strong to hold [57, 108]. We hope that our models and techniques for l_1-minimization dictionary learning can lead to better NMF identifiability conditions. On the other hand, we note that identifiability is only a minimal requirement for dictionary learning to be mathematically well-posed. This type of analysis does not give a practical algorithm to recover the dictionary. A number of recent works have been devoted to provable algorithms for sparse dictionary learning [83, 84, 85, 86, 87, 88]. For NMF, algorithms and theoretical analyses were provided for a convex formulation [109, 110]. We hope to extend these results to the more frequently used alternating minimization algorithm. A careful combination of the framework by [87] to analyze alternating minimization and the recent progress on nonnegative least squares [58] might help us to gain insight into this NMF algorithm.

Bibliography

[1] Jeremy A Miller et al. "Transcriptional landscape of the prenatal human brain". In: *Nature* 508.7495 (2014), pp. 199–206.

[2] Michael J Hawrylycz et al. "An anatomically comprehensive atlas of the adult human brain transcriptome". In: *Nature* 489.7416 (2012), pp. 391–399.

[3] Vanessa Almendro et al. "Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity". In: *Cell reports* 6.3 (2014), pp. 514–527.

[4] Vanessa Almendro et al. "Genetic and phenotypic diversity in breast tumor metastases". In: *Cancer research* 74.5 (2014), pp. 1338–1348.

[5] Philippe L Bedard et al. "Tumour heterogeneity in the clinic". In: *Nature* 501.7467 (2013), pp. 355–364.

[6] Marco Gerlinger et al. "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing". In: *New England Journal of Medicine* 366.10 (2012), pp. 883–892.

[7] Elza C de Bruin et al. "Spatial and temporal diversity in genomic instability processes defines lung cancer evolution". In: *Science* 346.6206 (2014), pp. 251–256.

[8] Jianjun Zhang et al. "Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing". In: *Science* 346.6206 (2014), pp. 256–259.

[9] Pavel Tomancak et al. "Systematic determination of patterns of gene expression during Drosophila embryogenesis". In: *Genome Biology* 3.12 (2002), research0088.1–88.14.

[10] Nir Yakoby et al. "A combinatorial code for pattern formation in drosophila oogenesis". In: *Developmental Cell* 15.5 (2008), pp. 725–737.

[11] Eric Lécuyer et al. "Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function". In: *Cell* 131.1 (2007), pp. 174–187.

[12] Arnim Jenett et al. "A GAL4-driver line resource for Drosophila neurobiology". In: *Cell reports* 2.4 (2012), pp. 991–1001.

[13] Aurélie Jory et al. "A survey of 6,300 genomic fragments for cis-regulatory activity in the imaginal discs of Drosophila melanogaster". In: *Cell reports* 2.4 (2012), pp. 1014–1024.

[14] Laurina Manning et al. "A resource for manipulating gene expression and analyzing cis-regulatory modules in the Drosophila CNS". In: *Cell reports* 2.4 (2012), pp. 1002–1013.

[15] Nicolas Pollet et al. "An atlas of differential gene expression during early Xenopus embryogenesis". In: *Mechanisms of development* 122.3 (2005), pp. 365–439.

[16] Kaoru S Imai et al. "Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: towards a comprehensive understanding of gene networks". In: *Development* 131.16 (2004), pp. 4047–4058.

[17] Constance M Smith et al. "The mouse gene expression database (GXD): 2007 update". In: *Nucleic acids research* 35.Database issue (2007), pp. D618–D623.

[18] Lorna Richardson et al. "EMAGE mouse embryo spatial gene expression database: 2010 update". In: *Nucleic acids research* 38.Database issue (2009), pp. D703–D709.

[19] Ed S Lein et al. "Genome-wide atlas of gene expression in the adult mouse brain". In: *Nature* 445.7124 (2007), pp. 168–176.

[20] Pavel Tomancak et al. "Global analysis of patterns of gene expression during Drosophila embryogenesis". In: *Genome biology* 8.7 (2007), R145.

[21] Ann S Hammonds et al. "Spatial expression of transcription factors in Drosophila embryonic organ development". In: *Genome Biology* 14.12 (2013), R140.

[22] Hanchuan Peng et al. "Automatic image analysis for gene expression patterns of fly embryos". In: *BMC Cell Biol* 8(Suppl 1): S7 (2007).

[23] Jie Zhou and Hanchuan Peng. "Automatic recognition and annotation of gene expression patterns of fly embryos". In: *Bioinformatics* 23.5 (2007), pp. 589–596.

[24] Iulian Pruteanu-Malinici, Daniel L Mace, and Uwe Ohler. "Automatic annotation of spatial expression patterns via sparse Bayesian factor models". In: *PLoS computational biology* 7.7 (2011), e1002098.

[25] Iulian Pruteanu-Malinici, William H Majoros, and Uwe Ohler. "Automated annotation of gene expression image sequences via non-parametric factor analysis and conditional random fields". In: *Bioinformatics* 29.13 (2013), pp. i27–i35.

[26] Lei Yuan et al. "Automated annotation of developmental stages of Drosophila embryos in images containing spatial patterns of expression". In: *Bioinformatics* 30.2 (2014), pp. 266–273.

[27] Erwin Frise, Ann S Hammonds, and Susan E Celniker. "Systematic image-driven analysis of the spatial Drosophila embryonic expression landscape". In: *Molecular systems biology* 6.1 (2010).

[28] Kriti Puniyani, Christos Faloutsos, and Eric P Xing. "SPEX2: automated concise extraction of spatial gene expression patterns from Fly embryo ISH images". In: *Bioinformatics* 26.12 (2010), pp. i47–i56.

[29] Daniel L Mace et al. "Extraction and comparison of gene expression patterns from 2D RNA in situ hybridization images". In: *Bioinformatics* 26.6 (2010), pp. 761–769.

[30] Kriti Puniyani and Eric P Xing. "GINI: from ISH images to gene interaction networks". In: *PLoS computational biology* 9.10 (2013), e1003227.

[31] Boris Adryan and Sarah A Teichmann. "The developmental expression dynamics of Drosophila melanogaster transcription factors". In: *Genome biology* 11.4 (2010), R40.

[32] Cordula Schulz and Diethard Tautz. "Autonomous concentration-dependent activation and repression of Kruppel by hunchback in the Drosophila embryo". In: *Development* 120.10 (1994), pp. 3043–3049.

[33] Evgeny Z Kvon et al. "Genome-scale functional characterization of Drosophila developmental enhancers in vivo". In: *Nature* 512.7512 (2014), pp. 91–95.

[34] Dusan Stanojevic, Stephen Small, and Michael Levine. "Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo". In: *Science* 254.5036 (1991), pp. 1385–1387.

[35] Margit Lohs-Schardin, Christoph Cremer, and Christiane Nüsslein-Volhard. "A fate map for the larval epidermis of Drosophila melanogaster: localized cuticle defects following irradiation of the blastoderm with an ultraviolet laser microbeam". In: *Developmental biology* 73.2 (1979), pp. 239–255.

[36] Volker Hartenstein. *Atlas of Drosophila Development*. New York: Plainview: Cold Spring Harbor Laboratory Press, 1993.

[37] Daniel D. Lee and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755 (1999), pp. 788–791.

[38] Bruno A Olshausen et al. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". In: *Nature* 381.6583 (1996), pp. 607–609.

[39] Yuan Wang et al. "Non-negative matrix factorization framework for face recognition". In: *International Journal of Pattern Recognition and Artificial Intelligence* 19.04 (2005), pp. 495–511.

[40] Wei Xu, Xin Liu, and Yihong Gong. "Document clustering based on non-negative matrix factorization". In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. SIGIR '03. Toronto, Canada: ACM, 2003, pp. 267–273. ISBN: 1-58113-646-3. DOI: 10.1145/860435.860485. URL: http://doi.acm.org/10.1145/860435.860485.

[41] Tuomas Virtanen. "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.3 (2007), pp. 1066–1074.

[42] Jean-Philippe Brunet et al. "Metagenes and molecular pattern discovery using matrix factorization". In: *Proceedings of the National Academy of Sciences* 101.12 (2004), pp. 4164–4169.

[43] Nicolas Gillis. "The why and how of nonnegative matrix factorization". In: *Regularization, Optimization, Kernels, and Support Vector Machines* 12 (2014), p. 257.

[44] Vincent YF Tan and Cédric Févotte. "Automatic relevance determination in nonnegative matrix factorization with the β-divergence". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35.7 (2013), pp. 1592–1605.

[45] Meng Sun and Xiongwei Zhang. "A stable approach for model order selection in nonnegative matrix factorization". In: *Pattern Recognition Letters* 54 (2015), pp. 97–102.

[46] Bin Yu. "Stability". In: *Bernoulli* 19.4 (2013), pp. 1484–1500.

[47] Julien Mairal et al. "Online learning for matrix factorization and sparse coding". In: *Journal of Machine Learning Research* 11 (2010), pp. 19–60.

[48] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 1 (1996), pp. 267–288.

[49] Christiane Nüsslein-Volhard and Eric Wieschaus. "Mutations affecting segment number and polarity in Drosophila". In: *Nature* 287.5785 (1980), pp. 795–801.

[50] Ch Nüsslein-Volhard, H Kluding, and G Jürgens. "Genes affecting the segmental subdivision of the Drosophila embryo". In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 50. Cold Spring Harbor Laboratory Press. 1985, pp. 145–154.

[51] Johannes Jaeger. "The gap gene network". In: *Cellular and Molecular Life Sciences* 68.2 (2011), pp. 243–274.

[52] Remi Gribonval and Karin Schnass. "Dictionary identification – sparse matrix-factorisation via l_1-minimisation". In: *IEEE Transactions on Information Theory* 56.7 (2010), pp. 3523–3539.

[53] Quan Geng, Huan Wang, and John Wright. "On the local correctness of L1 minimization for dictionary learning". In: *Technical report, preprint arXiv:1102.1249.2011* (2011).

[54] Remi Gribonval, Rodolphe Jenatton, and Francis Bach. "Sparse and spurious: dictionary learning with noise and outliers". In: *Technical report, preprint arXiv:1407.5155v3* (2014).

[55] Richard Weiszmann, Ann S Hammonds, and Susan E Celniker. "Determination of gene expression patterns using high-throughput RNA in situ hybridization to whole-mount Drosophila embryos". In: *Nature protocols* 4.5 (2009), pp. 605–618.

[56] Shunichi Amari, Andrzej Cichocki, and Howard Hua Yang. "A new learning algorithm for blind signal separation". In: *Advances in neural information processing systems* (1996), pp. 757–763.

[57] David Donoho and Victoria Stodden. "When does non-negative matrix factorization give a correct decomposition into parts?" In: *Advances in Neural Information Processing Systems 16*. Ed. by S. Thrun, L. K. Saul, and B. Schölkopf. MIT Press, 2004, pp. 1141–1148.

[58] Martin Slawski and Matthias Hein. "Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization". In: *Electronic Journal of Statistics* 7 (2013), pp. 3004–3056.

[59] Hanzhong Liu and Bin Yu. "Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression". In: *Electronic Journal of Statistics* 7 (2013), pp. 3124–3169.

[60] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Regularization paths for generalized linear Models via coordinate descent". In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. URL: `http://www.jstatsoft.org/v33/i01/`.

[61] Shovon I Ashraf et al. "The mesoderm determinant Snail collaborates with related zinc-finger proteins to control Drosophila neurogenesis". In: *The EMBO journal* 18.22 (1999), pp. 6426–6438.

[62] Aapo Hyvärinen. "Fast and robust fixed-point algorithms for independent component analysis". In: *IEEE Transactions on Neural Networks* 10.3 (1999), pp. 626–634.

[63] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[64] YX Rachel Wang and Haiyan Huang. "Review on statistical methods for gene network reconstruction using expression data". In: *Journal of theoretical biology* 362 (2014), pp. 53–61.

[65] Jonathan S Margolis et al. "Posterior stripe expression of hunchback is driven from two promoters by a common enhancer element". In: *Development* 121.9 (1995), pp. 3067–3077.

[66] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. "Atomic decomposition by basis pursuit". In: *SIAM Journal on Scientific Computing* 20.1 (1998), pp. 33–61.

[67] David L. Donoho and Michael Elad. "Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization". In: *Proceedings of the National Academy of Sciences of the United States of America* 100.5 (2003), pp. 2197–2202.

[68] J-J Fuchs. "On sparse representations in arbitrary redundant bases". In: *IEEE Transactions on Information Theory* 50.6 (2004), pp. 1341–1344.

[69] Emanuel Candes and Terrence Tao. "Decoding by linear programming". In: *IEEE Transactions on Information Theory* 51.12 (2005), pp. 4203–4215.

[70] S Mallat. *A Wavelet Tour of Signal Processing*. 3rd ed. Academic Press, 2008.

[71] Mark D Plumbley. "Dictionary learning for l_1-exact sparse coding". In: *Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 406–413.

[72] Michael Elad and Michal Aharon. "Image denoising via sparse and redundant representations over learned dictionaries". In: *Image Processing, IEEE Transactions on* 15.12 (2006), pp. 3736–3745.

[73] Gabriel Peyré. "Sparse modeling of textures". In: *Journal of Mathematical Imaging and Vision* 34.1 (2009), pp. 17–31.

[74] Roger Grosse et al. "Shift-invariance sparse coding for audio classification". In: *arXiv preprint arXiv:1206.5241* (2012).

[75] Michael Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer Science & Business Media, 2010.

[76] Ron Rubinstein, Alfred M. Bruckstein, and Michael Elad. "Dictionaries for sparse representation modeling". In: *Proceeds of the IEEE* 98.6 (2010), pp. 1045–1057.

[77] Julien Mairal, Francis Bach, and Jean Ponce. "Sparse modeling for image and vision processing". In: *arXiv preprint arXiv:1411.3230* (2014).

[78] Karin Schnass. "On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD". In: *Applied and Computational Harmonic Analysis* 37.3 (2014), pp. 464–491.

[79] Karin Schnass. "Local identification for overcomplete dictionaries". In: *Technical report, preprint ArXiv 1401.6354v1* (2015).

[80] Michal Aharon, Michael Elad, and Alfred M. Bruckstein. "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation". In: *Signal Processing, IEEE Transactions on* 54.11 (2006), pp. 4311–4322.

[81] Pando Georgiev, Fabian Theis, and Andrzej Cichocki. "Sparse component analysis and blind source separation of underdetermined mixtures". In: *Neural Networks, IEEE Transactions on* 16.4 (2005), pp. 992–996.

[82] Michal Aharon, Michael Elad, and Alfred M. Bruckstein. "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them". In: *Linear Algebra and its Applications* 416.1 (2006), pp. 48–67.

[83] Daniel A Spielman, Huan Wang, and John Wright. "Exact recovery of sparsely-used dictionaries". In: *arXiv preprint arXiv:1206.5882* (2012).

[84] Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. "Exact recovery of sparsely used overcomplete dictionaries". In: *arXiv preprint arXiv:1309.1952* (2014).

[85] S. Arora, Y. R. Ge, and A. Moitra. "New algorithms for learning incoherent and overcomplete dictionaries". In: *Technical report, preprint ArXiv 1308.6273* (2014).

[86] Alekh Agarwal et al. "Learning sparsely used overcomplete dictionaries via alternating minimization". In: *Technical report, preprint arXiv:1310.7991v2* (2014).

[87] Sanjeev Arora et al. "Simple, efficient, and neural algorithms for sparse coding". In: *arXiv preprint arXiv:1503.00778* (2015).

[88] Ju Sun, Qing Qu, and John Wright. "Complete dictionary recovery over the sphere". In: *arXiv preprint arXiv:1504.06785* (2015).

[89] Andreas Maurer and Massimiliano Pontil. "k-dimensional coding schemes in Hilbert spaces". In: *IEEE Transactions on Information Theory* 56.11 (2010), pp. 5839–5846.

[90] Daniel Vainsencher, Shie Mannor, and Alfred M Bruckstein. "The sample complexity of dictionary learning". In: *The Journal of Machine Learning Research* 12 (2011), pp. 3259–3281.

[91] Nishant A. Mehta and Alexander G. Gray. "On the sample complexity of predictive sparse coding". In: *CoRR* (2012).

[92] Rémi Gribonval et al. "Sample complexity of dictionary learning and other matrix factorizations". In: *arXiv preprint arXiv:1312.3790* (2013).

[93] Christopher Hillar and Friedrich T Sommer. "When can dictionary learning uniquely recover sparse data from subsamples?" In: *IEEE Transactions on Information Theory* 61.11 (2015), pp. 6290–6297.

[94] Pierre Comon. "Independent component analysis, a new concept?" In: *Signal processing* 36.3 (1994), pp. 287–314.

[95] Sanjeev Arora et al. "Provable ICA with unknown Gaussian noise, with implications for Gaussian mixtures and autoencoders". In: *Advances in Neural Information Processing Systems*. 2012, pp. 2375–2383.

[96] Sanjeev Arora et al. "Computing a nonnegative matrix factorization–provably". In: *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. ACM. 2012, pp. 145–162.

[97] Ben Recht et al. "Factoring nonnegative matrices with linear programs". In: *Advances in Neural Information Processing Systems*. 2012, pp. 1214–1222.

[98] P.A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[99] Alekh Agarwal et al. "Learning sparsely used overcomplete dictionaries". In: *JMLR: Workshop and Conference Proceedings* 35 (2014), pp. 1–15.

[100] Ming Yuan and Yi Lin. "Model selection and estimation in regression with grouped variables". In: *J.R.Statist.Soc.B* 68 (2006), pp. 49–57.

[101] José A Campos-Ortega and Volker Hartenstein. *The Embryonic Development of Drosophila Melanogaster*. Springer Science & Business Media, 2013.

[102] Judith A Lengyel and D David Iwaki. "It takes guts: the Drosophila hindgut as a model system for organogenesis". In: *Developmental biology* 243.1 (2002), pp. 1–19.

[103] Feng Yue et al. "A comparative encyclopedia of DNA elements in the mouse genome". In: *Nature* 515.7527 (2014), pp. 355–364.

[104] Erik Arner et al. "Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells". In: *Science* 347.6225 (2015), pp. 1010–1014.

[105] Indika Kahanda and Jennifer Neville. "Using transactional information to predict link strength in online social networks". In: *ICWSM* 1 (2009), pp. 74–81.

[106] David Liben-Nowell and Jon Kleinberg. "The link-prediction problem for social networks". In: *Journal of the American society for information science and technology* 58.7 (2007), pp. 1019–1031.

[107] Ryan N Lichtenwalter, Jake T Lussier, and Nitesh V Chawla. "New perspectives and methods in link prediction". In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2010, pp. 243–252.

[108] Kejun Huang, Nicholas Sidiropoulos, and Ananthram Swami. "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition". In: *Signal Processing, IEEE Transactions on* 62.1 (2014), pp. 211–224.

[109] Ben Recht et al. "Factoring nonnegative matrices with linear programs". In: *Advances in Neural Information Processing Systems*. 2012, pp. 1214–1222.

[110] Nicolas Gillis and Robert Luce. "Robust near-separable nonnegative matrix factorization using linear optimization". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 1249–1280.

[111] Peter Bühlmann and Sara Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.

Appendix A

Proofs of Part II results

In this appendix, we will present detailed proofs of our results on dictionary local identifiability. Let $L(\mathbf{D})$ be a function of $\mathbf{D} \in \mathcal{D}$ and $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$ be the collection of dictionaries $\mathbf{D}_t \in \mathcal{D}$ parameterized by $t \in \mathbb{R}$. By definition, $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$ passes through the reference dictionary $\mathbf{D}_0$ at $t = 0$. To ensure that $\mathbf{D}_0$ is a local minimum of $L(\mathbf{D})$, it suffices to have

$$\lim_{t\downarrow 0^+} \frac{L(\mathbf{D}_t) - L(\mathbf{D}_0)}{t} > 0 \text{ and } \lim_{t\uparrow 0^-} \frac{L(\mathbf{D}_t) - L(\mathbf{D}_0)}{t} < 0,$$

for all $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$ that is a smooth function of t. On the other hand, if either of the above strict inequalities holds in the reverse direction for some smooth $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$, then $\mathbf{D}_0$ is not a local minimum of $L(\mathbf{D})$.

Since $\mathbf{D}_0$ is full rank by assumption, the minimum eigenvalue of $\mathbf{M}_0 = \mathbf{D}_0^T\mathbf{D}_0$ is strictly greater than zero. By continuity of the minimum eigenvalue of $\mathbf{D}_t^T\mathbf{D}_t$ (see e.g. Bauer-Fike Thoerem), when $\mathbf{D}_t$ and $\mathbf{D}_0$ are sufficiently close, $\mathbf{D}_t$ should also be full rank. Thus without loss of generality we only need to work with full rank dictionary $\mathbf{D}_t$. For any full rank $\mathbf{D} \in \mathcal{D}$, there is a full rank matrix $\mathbf{A} \in \mathbb{R}^{K\times K}$ such that $\mathbf{D} = \mathbf{D}_0\mathbf{A}$. For any $k \in [\![K]\!]$, by the constraint $\|\mathbf{D}[, k]\|_2 = 1$, the matrix $\mathbf{A}$ should satisfy $\mathbf{A}[, k]^T\mathbf{M}_0\mathbf{A}[, k] = 1$. Define the set for all such $\mathbf{A}$'s as:

$$\mathcal{A} = \{\mathbf{A} \in \mathbb{R}^{K\times K} : \mathbf{A} \text{ is invertible and } \mathbf{A}[, k]^T\mathbf{M}_0\mathbf{A}[, k] = 1 \text{ for all } k \in [\![K]\!]\}. \tag{A.1}$$

It follows immediately that the set $\{\mathbf{D}_0\mathbf{A} : \mathbf{A} \in \mathcal{A}\}$ is the collection of $\mathbf{D} \in \mathcal{D}$ such that $\mathbf{D}$ is full rank. Thus, to ensure that $\mathbf{D}_0$ is a local minimum of $L(\mathbf{D})$, it suffices to show

$$\Delta^+(L, \{\mathbf{A}_t\}_t) \coloneqq \lim_{t\downarrow 0^+} \frac{L(\mathbf{D}_0\mathbf{A}_t) - L(\mathbf{D}_0)}{t} > 0, \tag{A.2}$$

$$\Delta^-(L, \{\mathbf{A}_t\}_t) \coloneqq \lim_{t\uparrow 0^-} \frac{L(\mathbf{D}_0\mathbf{A}_t) - L(\mathbf{D}_0)}{t} < 0, \tag{A.3}$$

for all smooth functions $\{\mathbf{A}_t\}_{t\in\mathbb{R}}$ with $\mathbf{A}_t \in \mathcal{A}$ and $\mathbf{A}_0 = \mathbf{I}$. In addition, to demonstrate that $\mathbf{D}_0$ is not a local minimum of $L(\mathbf{D})$, it suffices to have (A.2) or (A.3) hold in the

reverse direction for some $\{\mathbf{A}_t\}_t$ with the aforementioned properties. We will be using this characterization of local minimum to prove local identifiability results for both the population case and the finite sample case.

A.1 Proofs of the population results

Proof of Lemma 8.1

Proof. Since $\mathbb{E}||\mathbf{H}\boldsymbol{\alpha}_1||_1 = \sum_{j=1}^K \mathbb{E}|\mathbf{H}[j,]\boldsymbol{\alpha}_1|$, it suffices to compute $\mathbb{E}|\mathbf{H}[j,]\boldsymbol{\alpha}_1|$. Let S be any nonempty subset of $[\![K]\!]$. Recall that the random variable $\mathbf{S}_1 \subset [\![K]\!]$ denotes the support of random coefficient $\boldsymbol{\alpha}_1$. Conditioning on the event $\{\mathbf{S}_1 = S\}$, the random variable $\mathbf{H}[j,]\boldsymbol{\alpha}_1$ follows a normal distribution with mean 0 and standard deviation $\|\mathbf{H}[j,S]\|_2$. Hence

$$\mathbb{E}\,|\mathbf{H}[j,]\boldsymbol{\alpha}_1| = \mathbb{E}[\mathbb{E}[|\mathbf{H}[j,]\boldsymbol{\alpha}_1|\,|\mathbf{S}_1]] = \sqrt{\frac{2}{\pi}}\mathbb{E}\|\mathbf{H}[j,\mathbf{S}_1]\|_2.$$

(1) Under the s-sparse Gaussian model, $\mathbb{P}(\mathbf{S}_1 = S) = \binom{K}{s}^{-1}$ for any $|S| = s$. Thus we have

$$\mathbb{E}\|\mathbf{H}[j,\mathbf{S}_1]\|_2 = \binom{K}{s}^{-1} \sum_{S:|S|=s} \|\mathbf{H}[j,S]\|_2 = \frac{s}{K}|||\mathbf{H}[j,]|||_s.$$

Hence the objective function for the s-sparse Gaussian model is

$$L_{SG(s)}(\mathbf{D}) = \sum_{j=1}^K \mathbb{E}\,|\mathbf{H}[j,]\boldsymbol{\alpha}_1| = \sqrt{\frac{2}{\pi}}\frac{s}{K}\sum_{j=1}^K |||\mathbf{H}[j,]|||_s.$$

In particular, for $s = K$, $|||\mathbf{H}[j,]|||_K = \|\mathbf{H}[j,]\|_2$ and so

$$L_{SG(s)}(\mathbf{D}) = \sqrt{\frac{2}{\pi}}\sum_{j=1}^K \|\mathbf{H}[j,]\|_2.$$

(2) Under the Bernoulli(p)-Gaussian model, $\mathbb{P}(\mathbf{S}_1 = S) = p^{|S|}(1-p)^{K-|S|}$. So we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{H}[j,\mathbf{S}_1]\|_2] &= \sum_{k=1}^K \sum_{S:|S|=k} p^k(1-p)^{K-k}\|\mathbf{H}[j,S]\|_2 \\ &= p\sum_{k=0}^{K-1} \texttt{pbinom}(k;K-1,p)|||\mathbf{H}[j,]|||_{k+1}.\end{aligned}$$

Therefore for $p \in (0,1)$, the objective function under the Bernoulli-Gaussian model is

$$L_{BG(p)}(\mathbf{D}) = \sum_{j=1}^K \mathbb{E}\,|\mathbf{H}[j,]\boldsymbol{\alpha}_1| = \sqrt{\frac{2}{\pi}}p\sum_{j=1}^K |||\mathbf{H}[j,]|||_p.$$

Finally, if $p = 1$, we have

$$L_{BG(p)}(\mathbf{D}) = \sqrt{\frac{2}{\pi}} \sum_{j=1}^{K} \|\mathbf{H}[j,]\|_2.$$

□

Proof of Theorem 1

Proof. (1) Let us first consider the s-sparse Gaussian model. By (A.2) and (A.3), to ensure that $\mathbf{D}_0$ is a local minimum of $L_{SG(s)}(\mathbf{D})$, it suffices to show

$$\Delta^+(L_{SG(s)}, \{\mathbf{A}_t\}_t) > 0 \text{ and } \Delta^-(L_{SG(s)}, \{\mathbf{A}_t\}_t) < 0, \tag{A.4}$$

for all smooth functions $\{\mathbf{A}_t\}_t$ with $\mathbf{A}_t \in \mathcal{A}$, where $\mathcal{A}$ is defined in (A.1), and $\mathbf{A}_0 = \mathbf{I}$. Note that by Lemma 8.1,

$$\Delta^+(L_{SG(s)}, \{\mathbf{A}_t\}_t) = \sqrt{\frac{2}{\pi}} \frac{s}{K} \sum_{j=1}^{K} \lim_{t \downarrow 0^+} \frac{1}{t} \left(\left\| \left| \mathbf{A}_t^{-1}[j,] \right| \right\|_s - \left\| \left| \mathbf{I}[j,] \right| \right\|_s \right). \tag{A.5}$$

For a fixed $j \in [\![K]\!]$, we have

$$\binom{K-1}{s-1} \left\| \left| \mathbf{A}_t^{-1}[j,] \right| \right\|_s = \sum_{S:|S|=s, j \in S} \|\mathbf{A}_t^{-1}[j,S]\|_2 + \sum_{S:|S|=s, j \notin S} \|\mathbf{A}_t^{-1}[j,S]\|_2 \tag{A.6}$$

Denote by $\dot{\mathbf{A}}_0 \in \mathbb{R}^{K \times K}$ the derivative of $\{\mathbf{A}_t\}_t$ at $t = 0$. Since $\mathbf{A}_t \in \mathcal{A}$ for all $t \in \mathbb{R}$, it can be shown that

$$\mathbf{M}_0[,k]^T \dot{\mathbf{A}}_0[,k] = 0 \text{ for all } k \in [\![K]\!]. \tag{A.7}$$

By (A.7), we have

$$\dot{\mathbf{A}}_0[j,j] = -\sum_{i \neq j} \mathbf{M}_0[i,j] \dot{\mathbf{A}}_0[i,j] \text{ for all } j \in [\![K]\!]. \tag{A.8}$$

Now notice that

$$\left. \frac{d\mathbf{A}_t^{-1}}{dt} \right|_{t=0} = -\mathbf{A}_0^{-1} \dot{\mathbf{A}}_0 \mathbf{A}_0^{-1} = -\dot{\mathbf{A}}_0. \tag{A.9}$$

Combining the above equality with Lemma A.12 and A.13, we have

$$\lim_{t \downarrow 0^+} \frac{1}{t} \left(\|\mathbf{A}_t^{-1}[j,S]\|_2 - \|\mathbf{I}[j,S]\|_2 \right) = \begin{cases} -\dot{\mathbf{A}}_0[j,j] & \text{if } j \in S \\ \|\dot{\mathbf{A}}_0[j,S]\|_2 & \text{if } j \notin S \end{cases}$$

Therefore

$$\lim_{t\downarrow 0^+}\frac{1}{t}\left(\left|\left|\left|\mathbf{A}_t^{-1}[j,]\right|\right|\right|_s - |||\mathbf{I}[j,]|||_s\right) = -\dot{\mathbf{A}}_0[j,j] + \binom{K-1}{s-1}^{-1}\sum_{S:|S|=s,j\notin S}\|\dot{\mathbf{A}}_0[j,S]\|_2. \quad (A.10)$$

Combining (A.5), (A.6), (A.8) and (A.10), we have

$$\begin{aligned}\sqrt{\frac{\pi}{2}}\frac{K}{s}\Delta^+(L_{SG(s)},\{\mathbf{A}_t\}_t) &= -\sum_{j=1}^{K}\dot{\mathbf{A}}_0[j,j] + \binom{K-1}{s-1}^{-1}\sum_j\sum_{S:|S|=s,j\notin S}||\dot{\mathbf{A}}_0[j,S]||_2\\ &= \sum_{j=1}^{K}\Big(\sum_{i\neq j}\mathbf{M}_0[i,j]\dot{\mathbf{A}}_0[j,i] + \binom{K-1}{s-1}^{-1}\sum_{S:|S|=s,j\notin S}\|\dot{\mathbf{A}}_0[j,S]\|_2\Big).\end{aligned}$$

Similarly, one can show

$$\sqrt{\frac{\pi}{2}}\frac{K}{s}\Delta^-(L_{SG(s)},\{\mathbf{A}_t\}_t) = \sum_{j=1}^{K}\Big(\sum_{i\neq j}\mathbf{M}_0[i,j]\dot{\mathbf{A}}_0[j,i] - \binom{K-1}{s-1}^{-1}\sum_{S:|S|=s,j\notin S}\|\dot{\mathbf{A}}_0[j,S]\|_2\Big).$$

Thus for $s\in[\![K-1]\!]$, to establish (A.4) it suffices to require for each $j\in[\![K]\!]$,

$$\Big|\sum_{i\neq j}\mathbf{M}_0[i,j]\dot{\mathbf{A}}_0[j,i]\Big| < \frac{K-s}{K-1}\binom{K-2}{s-1}^{-1}\sum_{S:|S|=s,j\notin S}\|\dot{\mathbf{A}}_0[j,S]\|_2 = \frac{K-s}{K-1}\Big|\Big|\Big|\dot{\mathbf{A}}_0[j,-j]\Big|\Big|\Big|_s. \quad (A.11)$$

for any $\dot{\mathbf{A}}_0$ such that $\dot{\mathbf{A}}_0[j,-j]\neq 0$. Since $\dot{\mathbf{A}}_0[j,i]$ is a free variable for $i\neq j$, (A.11) is equivalent to

$$\Big|\mathbf{M}_0[-j,j]^T\mathbf{w}\Big| < \frac{K-s}{K-1},$$

for all $\mathbf{w}\in\mathbb{R}^{K-1}$ such that $|||\mathbf{w}|||_s = 1$. Thus by the definition of the dual norm, it suffices to have

$$|||\mathbf{M}_0[-j,j]|||_s^* = \sup_{|||\mathbf{w}|||_s=1}\Big|\mathbf{M}_0[-j,j]^T\mathbf{w}\Big| < \frac{K-s}{K-1}.$$

Therefore, the condition

$$\max_{1\le j\le K}|||\mathbf{M}_0[-j,j]|||_s^* < \frac{K-s}{K-1} = 1-\frac{s-1}{K-1}. \quad (A.12)$$

is sufficient for $\mathbf{D}_0$ to be locally identifiable with respect to the objective function $L_{SG(s)}$.

Similiarly, one can check that if the reversed strict inequality in (A.12) holds, $\mathbf{D}_0$ is not a local minimum of $L_{SG(s)}(\mathbf{D})$. Thus we complete the proof for the s-sparse model.

(2) Now consider the Bernoulli(p)-Gaussian model for $p \in (0,1)$. First of all, note that we have

$$\begin{aligned}
\sqrt{\frac{\pi}{2}}\frac{1}{p}\Delta^{\pm}(L_{BG(p)}, \{\mathbf{A}_t\}_t) &= \sum_{j=1}^{K} \lim_{t\to 0^{\pm}} \frac{1}{t}\left(\left|\!\left|\!\left| \mathbf{A}_t^{-1}[j,] \right|\!\right|\!\right|_p - \left|\!\left|\!\left| \mathbf{I}[j,] \right|\!\right|\!\right|_p \right) \\
&= \sum_{j=1}^{K} \left(\sum_{i\neq j} \mathbf{M}_0[i,j]\dot{\mathbf{A}}_0[j,i] \pm (1-p)\sum_{k=0}^{K-2} p^k (1-p)^{K-2-k} \sum_{S:|S|=k+1, j\notin S} \|\dot{\mathbf{A}}_0[j,S]\|_2 \right) \\
&= \sum_{j=1}^{K} \left(\dot{\mathbf{A}}_0[j,-j]^T \mathbf{M}_0[-j,j] \pm (1-p)\sum_{k=0}^{K-2} \texttt{pbinom}(k; K-2, p) \left|\!\left|\!\left| \dot{\mathbf{A}}_0[j,-j] \right|\!\right|\!\right|_{k+1} \right) \\
&= \sum_{j=1}^{K} \left(\dot{\mathbf{A}}_0[j,-j]^T \mathbf{M}_0[-j,j] \pm (1-p) \left|\!\left|\!\left| \dot{\mathbf{A}}_0[j,-j] \right|\!\right|\!\right|_p \right).
\end{aligned}$$

Thus, similar to the s-sparse Gaussian case, it can be shown that a sufficient condition for local identifiability is

$$\left| \mathbf{M}_0[-j,j]^T \mathbf{w} \right| < 1-p,$$

for all $j \in [\![K]\!]$ and all $\mathbf{w} \in \mathbb{R}^{K-1}$ such that $\left|\!\left|\!\left| \mathbf{w} \right|\!\right|\!\right|_p = 1$. The above condition is equivalent to

$$\max_{1\leq j\leq K} \left|\!\left|\!\left| \mathbf{M}_0[-j,j] \right|\!\right|\!\right|_p^* < 1-p.$$

The rest of the proof can be proceeded as in the case of the s-sparse Gaussian model.
(3) Now let us consider the non-sparse case where $s = K$ or $p = 1$. In this case, since the objective functions are the same under both models (see Theorem 1), we only need to consider the s-sparse Gaussian model. If $s = K$, the RHS quantity in Inequality (A.11) is zero. Thus, the reference dictionary is not locally identifiable if

$$\left| \mathbf{M}_0[-j,j]^T \mathbf{w} \right| > 0,$$

for some $j \in [\![K]\!]$ and $\mathbf{w} \in \mathbb{R}^{K-1}$. Thus, if $\mathbf{M}_0$ is not the identity matrix, or equivalently, if the reference dictionary $\mathbf{D}_0$ is not orthogonal, $\mathbf{D}_0$ is not locally identifiable.

Next, let us deal with the case where $\mathbf{D}_0$ is orthogonal. Let $\mathbf{D} \in \mathcal{D}$ be a full rank dictionary and $\mathbf{W} = \mathbf{D}^{-1}$. Since $\mathbf{D}_0$ is orthogonal, $\|\mathbf{W}[j,]\mathbf{D}_0\|_2 = \|\mathbf{W}[j,]\|_2$. By the fact that $\mathbf{W}\mathbf{D} = \mathbf{I}$ and $\|\mathbf{D}[,j]\|_2 = 1$, we have $1 = \mathbf{W}[j,]\mathbf{D}[,j] \leq \|\mathbf{W}[j,]\|_2 \|\mathbf{D}[,j]\|_2 = \|\mathbf{W}[j,]\|_2$, where the equality holds iff $\mathbf{W}[j,]^T = \pm\mathbf{D}[,j]$.

Under the K-sparse Gaussian model,

$$L_{SG(K)}(\mathbf{D}) = \sqrt{\frac{2}{\pi}} \sum_{j=1}^{K} \|\mathbf{W}[j,]\mathbf{D}_0\|_2 = \sqrt{\frac{2}{\pi}} \sum_{j=1}^{K} \|\mathbf{W}[j,]\|_2 \geq \sqrt{\frac{2}{\pi}} K = L_{SG(K)}(\mathbf{D}_0),$$

where the equality holds for any $\mathbf{D}$ such that $\mathbf{D}^T\mathbf{D} = \mathbf{I}$. Thus, $L_{SG(K)}(\mathbf{D}_0) = L_{SG(K)}(\mathbf{D}_0\mathbf{U})$ for any orthogonal matrix $\mathbf{U} \in \mathbb{R}^{K\times K}$, i.e. the objective function remains the same as we rotate $\mathbf{D}_0$. Therefore, $\mathbf{D}_0$ is not a local minimum of $L_{SG(K)}$.

In conclusion, $\mathbf{D}_0$ is not locally identifiable when $s = K$ or $p = 1$.

□

A.2 Proofs of the finite sample results: Theorem 2 and Theorem 3

Proof. We will first recall the signal generation procedure in Section 7.2. Let $\mathbf{z}$ be a K-dimensional standard Gaussian vector, and $\boldsymbol{\xi} \in \{0,1\}^K$ be either an s-sparse random vector or a Bernoulli random vector with probability p. Let $\mathbf{z}_1, ..., \mathbf{z}_N$ and $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_N$ be identical and independent copies of $\mathbf{z}$ and $\boldsymbol{\xi}$ respectively. For each $i \in [\![N]\!]$ and $j \in [\![K]\!]$, define $\boldsymbol{\alpha}_i[j] = \mathbf{z}_i[j]\boldsymbol{\xi}_i[j]$. For $S \subset [\![K]\!]$ with $1 \le |S| \le K-1$, define

$$\chi_i(S) = \begin{cases} 1 & \text{if } \boldsymbol{\xi}_i[k] = 1 \text{ for all } k \in S \text{ and } \boldsymbol{\xi}_i[k] = 0 \text{ for all } k \in S^c, \\ 0 & \text{otherwise.} \end{cases}$$

On the other hand, if $S = [\![K]\!]$, define $\chi_i(S) = 1$ if $\boldsymbol{\xi}_i[k] = 1$ for all $k \in [\![K]\!]$ and $\chi_i(S) = 0$ otherwise. As in the population case, in the following analysis we will work with full rank dictionaries $\mathbf{D}$. First of all, notice that

$$l(\mathbf{D}, \mathbf{x}_i) = \|\mathbf{D}^{-1}\mathbf{x}_i\|_1 = \|\mathbf{D}^{-1}\mathbf{D}_0\boldsymbol{\alpha}_i\|_1 = \sum_{j=1}^{K} \left|\mathbf{A}^{-1}[j,]\boldsymbol{\alpha}_i\right| = \sum_{j=1}^{K}\sum_{k=1}^{K}\left(\sum_{S:|S|=k} |\mathbf{A}^{-1}[j,S]\mathbf{z}_i[S]|\chi_i(S)\right).$$

Next, we have

$$\begin{aligned}
\Delta^+(l(.,\mathbf{x}_i), \{\mathbf{A}_t\}_t) &= \lim_{t\downarrow 0^+} \frac{1}{t}\left(l(\mathbf{D}_0\mathbf{A}_t, \mathbf{x}_i) - l(\mathbf{D}_0, \mathbf{x}_i)\right) \\
&= \sum_{j=1}^{K}\Bigg(-\sum_{k=1}^{K}\sum_{S:j\in S,|S|=k} \dot{\mathbf{A}}_0[j,j]|\mathbf{z}_i[j]|\chi_i(S) \\
&\quad - \mathbf{sgn}(\mathbf{z}_i[j])\sum_{k=2}^{K}\sum_{S:j\in S,|S|=k}\sum_{l\in S, l\neq j} \dot{\mathbf{A}}_0[j,l]\mathbf{z}_i[l]\chi_i(S) \\
&\quad + \sum_{k=1}^{K-1}\sum_{S:j\notin S,|S|=k} |\dot{\mathbf{A}}_0[j,S]\mathbf{z}_i[S]|\chi_i(S)\Bigg). \qquad \text{(A.13)}
\end{aligned}$$

Here $\mathbf{sgn}(x)$ is the sign function of $x \in \mathbb{R}$ such that $\mathbf{sgn}(x) = 1$ for $x > 0$, $\mathbf{sgn}(x) = -1$ for $x < 0$ and $\mathbf{sgn}(x) = 0$ for $x = 0$. By (A.8), the first term in (A.13) can be rearranged as

follows

$$-\sum_{j=1}^{K}|\mathbf{z}_i[j]|\sum_{k=1}^{K}\sum_{S:j\in S,|S|=k}\dot{\mathbf{A}}_0[j,j]\chi_i(S)=\sum_{j=1}^{K}|\mathbf{z}_i[j]|\sum_{k=1}^{K}\sum_{S:j\in S,|S|=k}\sum_{l\neq j}\mathbf{M}_0[l,j]\dot{\mathbf{A}}_0[l,j]\chi_i(S)$$
$$=\sum_{j=1}^{K}\sum_{l\neq j}\mathbf{M}_0[j,l]\dot{\mathbf{A}}_0[j,l]\left(|\mathbf{z}_i[l]|\sum_{k=1}^{K}\sum_{S:l\in S,|S|=k}\chi_i(S)\right).$$

The second term in (A.13) can be rewritten as

$$-\sum_{j=1}^{K}\mathbf{sgn}(\mathbf{z}_i[j])\times\sum_{l\neq j}(\dot{\mathbf{A}}_0[j,l]\mathbf{z}_i[l])\times\sum_{k=2}^{K}\sum_{S:\{j,l\}\in S,|S|=k}\chi_i(S).$$

For $j,l\in[\![K]\!]$ such that $j\neq l$, define the following quantities

$$\mathbf{F}_i[l,j]=\mathbf{M}_0[j,l]|\mathbf{z}_i[l]|\sum_{k=1}^{K}\sum_{S:l\in S,|S|=k}\chi_i(S), \tag{A.14}$$

$$\mathbf{G}_i[l,j]=\mathbf{sgn}(\mathbf{z}_i[j])\mathbf{z}_i[l]\sum_{k=2}^{K}\sum_{S:\{j,l\}\in S,|S|=s}\chi_i(S), \tag{A.15}$$

whereas $\mathbf{F}[j,j]=\mathbf{G}[j,j]=0$. For each $j\in[\![K]\!]$, also define

$$\mathbf{t}_i[j](\mathbf{w})=\sum_{k=1}^{K-1}\sum_{S:j\notin S,|S|=k}|\mathbf{w}[S]^T\mathbf{z}_i[S]|\chi_i(S). \tag{A.16}$$

Let $\bar{\mathbf{F}}$, $\bar{\mathbf{G}}$ and $\bar{\mathbf{t}}$ be the sample average of $\mathbf{F}_i$, $\mathbf{G}_i$ and $\mathbf{t}_i$ respectively. With the definitions (A.14) – (A.16), we have

$$\begin{aligned}\Delta^+(L_N,\{\mathbf{A}_t\}_t)&=\frac{1}{N}\sum_{i=1}^{N}\Delta^+(l(.,\mathbf{x}_i),\{\mathbf{A}_t\}_t)\\&=\sum_{j=1}^{K}\frac{1}{N}\sum_{i=1}^{N}\left(\dot{\mathbf{A}}_0[j,]\mathbf{F}_i[,j]+\dot{\mathbf{A}}_0[j,]\mathbf{G}_i[,j]+\mathbf{t}_i[j](\dot{\mathbf{A}}_0[j,])\right)\\&=\sum_{j=1}^{K}\left(\dot{\mathbf{A}}_0[j,]\bar{\mathbf{F}}[,j]-\dot{\mathbf{A}}_0[j,]\bar{\mathbf{G}}[,j]+\bar{\mathbf{t}}[j](\dot{\mathbf{A}}_0[j,])\right)\end{aligned}$$

On the other hand,

$$\Delta^-(L_N,\{\mathbf{A}_t\}_t)=\sum_{j=1}^{K}\left(\dot{\mathbf{A}}_0[j,]\bar{\mathbf{F}}[,j]-\dot{\mathbf{A}}_0[j,]\bar{\mathbf{G}}[,j]-\bar{\mathbf{t}}[j](\dot{\mathbf{A}}_0[j,])\right).$$

Now for $j \in [\![K]\!]$, $s \in [\![K-1]\!]$ and $p \in (0,1)$, define

$$\mathcal{E}_j(s) = \{\mathbf{w} \in \mathbb{R}^K, |||\mathbf{w}[-j]|||_s = 1, \mathbf{w}[j] = 0\},$$

$$\mathcal{F}_j(p) = \{\mathbf{w} \in \mathbb{R}^K, |||\mathbf{w}[-j]|||_p = 1, \mathbf{w}[j] = 0\}.$$

Thus to ensure that $\mathbf{D}_0$ is a local minimum, it suffices to have for each $j \in [\![K]\!]$,

$$H_j(\mathbf{w}) \coloneqq \left|\mathbf{w}^T \bar{\mathbf{F}}[,j] - \mathbf{w}^T \bar{\mathbf{G}}[,j]\right| - \bar{\mathbf{t}}[j](\mathbf{w}) < 0,$$

for all $\mathbf{w} \in \mathcal{E}_j(s)$ for the s-sparse Gaussian model or all $\mathbf{w} \in \mathcal{F}_j(p)$ for the Bernoulli(p)-Gaussian model.

(1) For the s-sparse Gaussian model, let $j \in [\![K]\!]$ and define

$$h_j(\mathbf{w}) = \sqrt{\frac{2}{\pi}} \frac{s}{K} \left(\left|\mathbf{w}^T \mathbf{M}_0[,j]\right| - \frac{K-s}{K-1} \right),$$

which can be thought of as the expected value of $H_j(\mathbf{w})$. Note that by triangle inequality,

$$\begin{aligned}
&\sup_{\mathbf{w} \in \mathcal{E}_j(s)} |H_j(\mathbf{w}) - h_j(\mathbf{w})| \\
\leq &\sup_{\mathbf{w} \in \mathcal{E}_j(s)} \left| \mathbf{w}^T \left(\bar{\mathbf{F}}[,j] - \sqrt{\frac{2}{\pi}} \frac{s}{K} \mathbf{M}_0[,j] \right) \right| + \sup_{\mathbf{w} \in \mathcal{E}_j(s)} \left|\mathbf{w}^T \bar{\mathbf{G}}[,j]\right| + \sup_{\mathbf{w} \in \mathcal{E}_j(s)} \left| \bar{\mathbf{t}}[j](\mathbf{w}) - \sqrt{\frac{2}{\pi}} \frac{s}{K} \frac{K-s}{K-1} \right| \\
= &\left|\left|\left| \bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}} \frac{s}{K} \mathbf{M}_0[-j,j] \right|\right|\right|_s^* + \left|||\bar{\mathbf{G}}[-j,j]|||_s^* + \sup_{\mathbf{w} \in \mathcal{E}_j(s)} \left| \bar{\mathbf{t}}[j](\mathbf{w}) - \sqrt{\frac{2}{\pi}} \frac{s}{K} \frac{K-s}{K-1} \right|.
\end{aligned} \tag{A.17}$$

Thus, $\sup_{\mathbf{w} \in \mathcal{E}_j(s)} |H_j(\mathbf{w}) - h_j(\mathbf{w})| > \frac{s}{K}\epsilon$ implies at least one of the three terms on the RHS is greater than $\frac{s}{K}\frac{\epsilon}{3}$. Using a union bound and by Lemma A.1–A.3, we have

$$\begin{aligned}
\mathbb{P}\left\{ \sup_{\mathbf{w} \in \mathcal{E}_j(s)} |H_j(\mathbf{w}) - h_j(\mathbf{w})| > \frac{s}{K}\epsilon \right\} \leq &\, 2K \exp\left(-\frac{N\epsilon^2}{108K \|\mathbf{M}_0[-j,j]\|_\infty} \right) \\
&+ 2K \exp\left(-\frac{s}{K} \frac{N\epsilon^2}{18(s/K)s + 9\sqrt{2s}} \right) \\
&+ 3\left(\frac{24K}{\epsilon s} + 1 \right)^K \exp\left(-\frac{s}{K} \frac{N\epsilon^2}{360} \right).
\end{aligned} \tag{A.18}$$

It is easy to see that the event $\left\{\sup_{\mathbf{w} \in \mathcal{E}_j(s)} |H_j(\mathbf{w}) - h_j(\mathbf{w})| \leq \frac{s}{K}\epsilon\right\}$ implies

$$\sup_{\mathbf{w} \in \mathcal{E}_j(s)} h_j(\mathbf{w}) - \frac{s}{K}\epsilon \leq \sup_{\mathbf{w} \in \mathcal{E}_j(s)} H_j(\mathbf{w}) \leq \sup_{\mathbf{w} \in \mathcal{E}_j(s)} h_j(\mathbf{w}) + \frac{s}{K}\epsilon. \tag{A.19}$$

On the other hand,

$$\sup_{\mathbf{w}\in\mathcal{E}_j(s)} h_j(\mathbf{w}) = \sqrt{\frac{2}{\pi}}\frac{s}{K}\left(|||\mathbf{M}_0[-j,j]|||_s^* - \frac{K-s}{K-1}\right).$$

Thus, if $|||\mathbf{M}_0[-j,j]|||_s^* < \frac{K-s}{K-1} - \sqrt{\frac{\pi}{2}}\epsilon$, $\sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) < 0$ except with probability at most the bound in (A.18). To ensure $\mathbf{D}_0$ to be a local minimum, it suffices to have $\sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) < 0$ for all $j \in [\![K]\!]$. Thus, if $|||\mathbf{M}_0[-j,j]|||_s^* < \frac{K-s}{K-1} - \sqrt{\frac{\pi}{2}}\epsilon$ for all $j \in [\![K]\!]$, we have

$$\begin{aligned}
\mathbb{P}\left\{\mathbf{D}_0 \text{ is locally identifiable}\right\} &\geq \mathbb{P}\left\{\max_j \sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) < 0\right\} \\
&\geq 1 - \mathbb{P}\left\{\max_j \sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) \geq 0\right\} \\
&\geq 1 - \sum_{j=1}^{K} \mathbb{P}\left\{\sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) \geq 0\right\} \\
&\geq 1 - \sum_{j=1}^{K} \mathbb{P}\left\{\sup_{\mathbf{w}\in\mathcal{E}_j(s)} |H_j(\mathbf{w}) - h_j(\mathbf{w})| > \frac{s}{K}\epsilon\right\} \\
&\geq 1 - 2K^2 \exp\left(-\frac{N\epsilon^2}{108K \max_{l\neq j} |\mathbf{M}_0[l,j]|}\right) \\
&\quad - 2K^2 \exp\left(-\frac{s}{K}\frac{N\epsilon^2}{18(s/K)s + 9\sqrt{2s}}\right) \\
&\quad - 3K\left(\frac{24K}{\epsilon s} + 1\right)^K \exp\left(-\frac{s}{K}\frac{N\epsilon^2}{360}\right).
\end{aligned}$$

On the other hand, to ensure $\mathbf{D}_0$ is not locally identifiable with high probability, it suffices to have $|||\mathbf{M}_0[-j,j]|||_s^* > \frac{K-s}{K-1} + \sqrt{\frac{\pi}{2}}\epsilon$ for some $j \in [\![K]\!]$. Indeed, under that condition, the

LHS inequality in (A.19) implies $\sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) > 0$. Therefore

$$\begin{aligned}
\mathbb{P}\left\{\mathbf{D}_0 \text{ is not locally identifiable}\right\} &\geq \mathbb{P}\left\{\sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) > 0\right\} \\
&\geq 1 - \mathbb{P}\left\{\sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) \leq 0\right\} \\
&\geq 1 - \mathbb{P}\left\{\sup_{\mathbf{w}\in\mathcal{E}_j(s)} |H_j(\mathbf{w}) - h_j(\mathbf{w})| > \frac{s}{K}\epsilon\right\} \\
&\geq 1 - 2K\exp\left(-\frac{N\epsilon^2}{108K\|\mathbf{M}_0[-j,j]\|_\infty}\right) \\
&\quad - 2K\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{18(s/K)s + 9\sqrt{2s}}\right) \\
&\quad - 3\left(\frac{24K}{\epsilon s} + 1\right)^K \exp\left(-\frac{s}{K}\frac{N\epsilon^2}{360}\right).
\end{aligned}$$

(2) For the Bernoulli(p)-Gaussian model, define

$$\nu_j(\mathbf{w}) = \sqrt{\frac{2}{\pi}} p\left(\left|\mathbf{w}^T\mathbf{M}_0[,j]\right| - (1-p)\right).$$

Similar to (A.17), by triangle inequality,

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{F}_j(p)} |H_j(\mathbf{w}) - \nu_j(\mathbf{w})| \\
\leq &\left|\left|\left|\bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}} p\mathbf{M}_0[-j,j]\right|\right|\right|_p^* + \left|\left|\left|\bar{\mathbf{G}}[-j,j]\right|\right|\right|_p^* + \sup_{\mathbf{w}\in\mathcal{F}_j(p)} \left|\bar{\mathbf{t}}[j](\mathbf{w}) - \sqrt{\frac{2}{\pi}} p(1-p)\right|.
\end{aligned}$$

Then the analysis can be carried out in a similar manner using the parallel version of the concentration inequalities, i.e. Part 2 of Lemma A.1–A.3. □

A.3 Concentration inequalities

We will make frequent use of the following version of Bernstein's inequality. The proof of the inequality can be found in, e.g. Chapter 14 of [111].

Theorem A.3.1. *(Bernstein's inequality) Let $Y_1, ..., Y_N$ be independent random variables that satisfy the moment condition*

$$\mathbb{E}Y_i^m \leq \frac{1}{2} \times V \times m! \times B^{m-2},$$

for integers $m \geq 2$. Then

$$\mathbb{P}\left\{\frac{1}{N}|\sum_{i=1}^{N} Y_i - \mathbb{E}Y_i| > \epsilon\right\} \leq 2\exp\left(-\frac{N\epsilon^2}{2V + 2B\epsilon}\right).$$

Lemma A.1. *(Uniform concentration of $\bar{\mathbf{F}}[-j,j]$) For $i \in [\![N]\!]$, let $\mathbf{F}_i \in \mathbb{R}^{K\times K}$ be defined as in (A.14) and $\bar{\mathbf{F}} = (1/N)\sum_{i=1}^{N}\mathbf{F}_i$.*

1. *Under the s-sparse Gaussian model with $s \in [\![K-1]\!]$,*

$$\mathbb{P}\left\{\left|\left|\left|\bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}}\frac{s}{K}\mathbf{M}_0[-j,j]\right|\right|\right|_s^* > \frac{s}{K}\epsilon\right\} \leq 2K\exp\left(-\frac{N\epsilon^2}{12K\|\mathbf{M}_0[-j,j]\|_\infty}\right),$$

   for $0 < \epsilon \leq 1$.

2. *Under the Bernoulli-Gaussian model with parameter $p \in (0,1)$,*

$$\mathbb{P}\left\{\left|\left|\left|\bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}}p\mathbf{M}_0[-j,j]\right|\right|\right|_p^* > p\epsilon\right\} \leq 2K\exp\left(-\frac{N\epsilon^2}{12(K+2p^{-1})\|\mathbf{M}_0[-j,j]\|_\infty}\right),$$

   for $0 < \epsilon \leq 1$.

In particular, if $\|\mathbf{M}_0[-j,j]\|_\infty = 0$, then the RHS bound is trivially zero.

Proof. (1) First of all, we will prove the inequality for the s-sparse model. Notice that by Lemma 8.2, we have

$$\begin{aligned}\left|\left|\left|\bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}}\frac{s}{K}\mathbf{M}_0[-j,j]\right|\right|\right|_s^* &\leq \max_{|S|=s, j\notin S}\|\bar{\mathbf{F}}[S,j] - \sqrt{\frac{2}{\pi}}\frac{s}{K}\mathbf{M}_0[S,j]\|_2 \\ &\leq \sqrt{s}\max_{l\neq j}|\bar{\mathbf{F}}[l,j] - \sqrt{\frac{2}{\pi}}\frac{s}{K}\mathbf{M}_0[l,j]|.\end{aligned}$$

For convenience, define

$$\mathbf{v}_i[l] = |\mathbf{z}_i[l]|\sum_{k=1}^{K}\sum_{|S|=k, l\in S}\chi_i(S) - \sqrt{\frac{2}{\pi}}\frac{s}{K}.$$

for $i \in [\![N]\!]$ and $l \in [\![K]\!]$. Note that $\sum_{k=1}^{K}\sum_{l\in S,|S|=k}\chi_i(S) = 1$ with probability $\binom{K}{s}^{-1}\binom{K-1}{s-1} = \frac{s}{K}$. Thus

$$\mathbb{E}\left(\sum_{k=1}^{K}\sum_{|S|=k,l\in S}\chi_i(S)\right)^m = \frac{s}{K}.$$

For $m \geq 1$, by Jensen's inequality $|\frac{a+b}{2}|^m \leq \frac{1}{2}(|a|^m + |b|^m)$ and $\mathbb{E}|Z|^m \geq (\mathbb{E}|Z|)^m = \left(\frac{2}{\pi}\right)^{\frac{m}{2}}$, where Z is a standard Gaussian variable. In addition, $\mathbb{E}|Z|^m \leq (m-1)!! \leq 2^{-\frac{m}{2}} m!$. Hence

$$\begin{aligned}
\mathbb{E}|\mathbf{v}_i[l]|^m &\leq 2^{m-1}\left(\mathbb{E}|\mathbf{z}_i[l]|^m + \left(\frac{2}{\pi}\right)^{\frac{m}{2}}\left(\frac{s}{K}\right)^m\right) \\
&\leq 2 \times \mathbb{E}|Z|^m \times 2^{m-1} \\
&\leq 2 \times \left(\frac{1}{2}\right)^{\frac{m}{2}} m! \times 2^{m-1} \\
&= \frac{1}{2} \times \frac{4s}{K} \times m! \times (\sqrt{2})^{m-2}.
\end{aligned}$$

Thus by Bernstein's inequality, we have

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^N \mathbf{v}_i[l]\right| > \epsilon\right\} \leq 2\exp\left(-\frac{N\epsilon^2}{2(4\frac{s}{K} + \sqrt{2}\epsilon)}\right).$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left\{\left|\mathbf{M}_0[j,l]\frac{1}{N}\sum_{i=1}^N \mathbf{v}_i[l]\right| > \frac{s}{K}\epsilon\right\} &\leq 2\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{2(4\mathbf{M}_0[j,l]^2 + \sqrt{2}\,|\mathbf{M}_0[j,l]|\,\epsilon)}\right) \\
&\leq 2\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{2\,|\mathbf{M}_0[j,l]|\,(4+\sqrt{2}\epsilon)}\right) \\
&\leq 2\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{12\,|\mathbf{M}_0[j,l]|}\right).
\end{aligned}$$

for $\epsilon \leq 1$. Notice that if $\mathbf{M}_0[j,l] = 0$ the LHS probability is trivially zero. Using a union bound, we have

$$\begin{aligned}
\mathbb{P}\left\{\|\bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}}\frac{s}{K}\mathbf{M}_0[-j,j]\|_\infty > \frac{s}{K}\epsilon\right\} &= \mathbb{P}\left\{\max_{l\neq j}|\mathbf{M}_0[j,l]\frac{1}{N}\sum_{i=1}^N \mathbf{v}_i[l]| > \epsilon\right\} \\
&\leq 2K\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{12\|\mathbf{M}_0[-j,j]\|_\infty}\right).
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{P}\left\{\left|\left|\left|\bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}}\frac{s}{K}\mathbf{M}_0[-j,j]\right|\right|\right|_s^* > \frac{s}{K}\epsilon\right\} &\leq \mathbb{P}\left\{\sqrt{s}\|\bar{\mathbf{F}}[,j] - \sqrt{\frac{2}{\pi}}\frac{s}{K}\mathbf{M}_0[,j]\|_\infty > \frac{s}{K}\epsilon\right\} \\
&\leq 2K\exp\left(-\frac{N\epsilon^2}{12K\|\mathbf{M}_0[-j,j]\|_\infty}\right).
\end{aligned}$$

(2) Now let us consider the Bernoulli-Gaussian model. Notice that by Lemma A.6, for $\frac{s-1}{K-1} \geq p$, we have

$$\left|\!\left|\!\left|\bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}} p\mathbf{M}_0[-j,j]\right|\!\right|\!\right|_p^* \leq \left|\!\left|\!\left|\bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}} p\mathbf{M}_0[-j,j]\right|\!\right|\!\right|_s^* \leq \sqrt{s}\left\|\bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}} p\mathbf{M}_0[-j,j]\right\|_\infty.$$

Now let $s = \lceil pK - p + 1\rceil \leq pK + 2$. For $i \in [\![N]\!]$ and $l \in [\![K]\!]$, define

$$\mathbf{u}_i[l] = |\mathbf{z}_i[l]| \sum_{k=1}^{K} \sum_{|S|=s, l\in S} \chi_i(S) - \sqrt{\frac{2}{\pi}} p.$$

Note that the event $\left\{\sum_{k=1}^{K}\sum_{|S|=k,l\in S}\chi_i(S) = 1\right\}$ is the same as the event that $\{\boldsymbol{\alpha}_i[l] = 1\}$, which, happens with probability p. Thus

$$\mathbb{E}\left(\sum_{k=1}^{K}\sum_{|S|=k,l\in S}\chi_i(S)\right)^m = p.$$

Similar to the case of s-sparse model,

$$\mathbb{E}\,|\mathbf{u}_i[l]|^m \leq \frac{1}{2} \times 4p \times m! \times \left(\sqrt{2}\right)^{m-2}.$$

By Bernstein's inequality, we have

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}\mathbf{u}_i[l]\right| > \epsilon\right\} \leq 2\exp\left(-\frac{N\epsilon^2}{2(4p+\sqrt{2}\epsilon)}\right).$$

Therefore

$$\begin{aligned}\mathbb{P}\left\{\left|\!\left|\!\left|\bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}} p\mathbf{M}_0[-j,j]\right|\!\right|\!\right|_p^* > p\epsilon\right\} &\leq \mathbb{P}\left\{\sqrt{s}\|\bar{\mathbf{F}}[,j] - \sqrt{\frac{2}{\pi}} p\mathbf{M}_0[,j]\|_\infty > p\epsilon\right\} \\ &\leq 2K\exp\left(-\frac{p}{s}\frac{N\epsilon^2}{2\|\mathbf{M}_0[-j,j]\|_\infty(4+\sqrt{2}\epsilon)}\right) \\ &\leq 2K\exp\left(-\frac{N\epsilon^2}{12(K+2p^{-1})\|\mathbf{M}_0[-j,j]\|_\infty}\right),\end{aligned}$$

for $\epsilon \leq 1$. □

Lemma A.2. *(Uniform concentration of $\bar{\mathbf{G}}[-j,j]$) For $i \in [\![N]\!]$, let $\mathbf{G}_i \in \mathbb{R}^{K\times K}$ be defined as in (A.15) and $\bar{\mathbf{G}} = (1/N)\sum_{i=1}^N \mathbf{G}_i$.*

1. *Under the s-sparse Gaussian model with $s \in [\![K-1]\!]$,*

$$\mathbb{P}\left\{ \left|\!\left|\!\left| \bar{\mathbf{G}}[-j,j] \right|\!\right|\!\right|_s^* > \frac{s}{K}\epsilon \right\} \le 2K \exp\left(-\frac{s}{K} \frac{N\epsilon^2}{2(s/K)s + \sqrt{2s}} \right),$$

   for $0 < \epsilon \le 1$.

2. *Under the Bernoulli-Gaussian model with parameter $p \in (0,1)$,*

$$\mathbb{P}\left\{ \left|\!\left|\!\left| \bar{\mathbf{G}}[-j,j] \right|\!\right|\!\right|_p^* > p\epsilon \right\} \le 2K \exp\left(-p \frac{N\epsilon^2}{p(pK+2) + \sqrt{2(pK+2)}} \right),$$

   for $0 < \epsilon \le 1$.

Proof. The proof is highly similar to that of Lemma A.1 and so we will omit some common steps.
(1) We first prove the concentration inequality for the s-sparse model. Notice that

$$\left|\!\left|\!\left| \bar{\mathbf{G}}[-j,j] \right|\!\right|\!\right|_s^* \le \sqrt{s} \max_{l\neq j} |\bar{\mathbf{G}}[l,j]|.$$

In addition,

$$\begin{aligned}\mathbb{E}\left(\sum_{k=2}^K \sum_{\{j,l\}\in S, |S|=k} \chi_i(S) \right)^m &= \mathbb{E}\left(\sum_{k=2}^K \sum_{|S|=k, \{j,l\}\in S} \chi_i(S) \right) \\ &= \binom{K}{s}^{-1} \binom{K-2}{s-2} = \frac{s(s-1)}{K(K-1)} \le (\frac{s}{K})^2.\end{aligned}$$

Thus

$$\mathbb{E}|\mathbf{G}_i[l,j]|^m \le 2^{-m/2} m! \times (\frac{s}{K})^2 = \frac{1}{2} \times (\frac{s}{K})^2 \times m! \times (\frac{1}{\sqrt{2}})^{m-2}.$$

By Bernstein inequality:

$$\mathbb{P}\left\{ |\frac{1}{N}\sum_{i=1}^N \mathbf{G}_i[l,j]| > \epsilon \right\} \le 2\exp\left(-\frac{N\epsilon^2}{2(s/K)^2 + \sqrt{2}\epsilon)} \right).$$

Thus we have

$$\begin{aligned}\mathbb{P}\left\{\left|\left|\left|\bar{\mathbf{G}}[-j,j]\right|\right|\right|_s^* > \frac{s}{K}\epsilon\right\} &\le \mathbb{P}\left\{\sqrt{s}\max_{l\neq j}|\bar{\mathbf{G}}[l,j]| > \frac{s}{K}\epsilon\right\}\\ &\le 2K\exp\left(-\frac{(s/K)^2N(\epsilon^2/s)}{2(s/K)^2+\sqrt{2}(s/K)(\epsilon/\sqrt{s}))}\right)\\ &\le 2K\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{2(s/K)s+\sqrt{2s}\epsilon}\right)\\ &\le 2K\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{2(s/K)s+\sqrt{2s}}\right),\end{aligned}$$

for $\epsilon \le 1$.
(2) For Bernoulli-Gaussian model, notice that

$$\left|\left|\left|\bar{\mathbf{G}}[-j,j]\right|\right|\right|_p^* \le \left|\left|\left|\bar{\mathbf{G}}[-j,j]\right|\right|\right|_s^* \le \sqrt{s}\max_{l\neq j}|\bar{\mathbf{G}}[l,j]|,$$

for $s = \lceil pK - p + 1\rceil \le pK + 2$. Also,

$$\mathbb{E}|\mathbf{G}_i[l,j]|^m \le 2^{-m/2}m! \times p^2 = \frac{1}{2}\times p^2 \times m! \times (\frac{1}{\sqrt{2}})^{m-2}.$$

Thus

$$\begin{aligned}\mathbb{P}\left\{\left|\left|\left|\bar{\mathbf{G}}[-j,j]\right|\right|\right|_s^* > p\epsilon\right\} &\le \mathbb{P}\left\{\sqrt{s}\max_{l\neq j}|\bar{\mathbf{G}}[l,j]| > p\epsilon\right\}\\ &\le 2K\exp(-p\frac{N(\epsilon^2)}{2ps+\sqrt{2s}})\\ &\le 2K\exp\left(-p\frac{N\epsilon^2}{p(pK+2)+\sqrt{2(pK+2)}}\right),\end{aligned}$$

for $\epsilon \le 1$. □

Lemma A.3. *(Uniform concentration of $\bar{\mathbf{t}}[j](\mathbf{w})$) For $i \in [\![N]\!]$, let $\mathbf{t}_i$ be a function from $\mathbb{R}^K$ to $\mathbb{R}^K$ defined as in (A.16) and $\bar{\mathbf{t}} = (1/N)\sum_{i=1}^N \mathbf{t}_i$. Recall that for $j \in [\![K]\!]$, $s \in [\![K-1]\!]$ and $p \in (0,1)$,*

$$\begin{aligned}\mathcal{E}_j(s) &= \{\mathbf{w}\in\mathbb{R}^K, \left|\left|\left|\mathbf{w}[-j]\right|\right|\right|_s = 1, \mathbf{w}[j] = 0\},\\ \mathcal{F}_j(p) &= \{\mathbf{w}\in\mathbb{R}^K, \left|\left|\left|\mathbf{w}[-j]\right|\right|\right|_p = 1, \mathbf{w}[j] = 0\}.\end{aligned}$$

1. *Under the s-sparse Gaussian model with $s \in [\![K-1]\!]$,*

$$\mathbb{P}\left\{\sup_{w\in\mathcal{E}_j(s)}|\bar{\mathbf{t}}[j](\mathbf{w}) - \sqrt{\frac{2}{\pi}}\frac{s}{K}\frac{K-s}{K-1}| > \frac{s}{K}\epsilon\right\} \le 3\left(\frac{8K}{\epsilon s}+1\right)^K\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{40}\right),$$

for $0 < \epsilon \le \frac{1}{2}$.

2. Under the Bernoulli-Gaussian model with parameter $p \in (0,1)$,

$$\mathbb{P}\left\{\sup_{w\in\mathcal{F}_j(p)} |\bar{\mathbf{t}}[j](\mathbf{w}) - \sqrt{\frac{2}{\pi}}p(1-p)| > p\epsilon\right\} \leq 3\left(\frac{8}{\epsilon p}+1\right)^K \exp\left(-p\frac{N\epsilon^2}{40}\right),$$

for $0 < \epsilon \leq \frac{1}{2}$.

Proof. (1) Under the s-sparse model, we have

$$\begin{aligned}\mathbb{E}|\mathbf{t}_i[j](\mathbf{w})|^m &= \mathbb{E}\left(\sum_{|S|=s, j\notin S} |\mathbf{w}[S]^T\mathbf{z}_i[S]|\chi_i(S)\right)^m \\ &= \sum_{|S|=s,j\notin S} \mathbb{E}|\mathbf{w}[S]^T\mathbf{z}_i[S]|^m \mathbb{E}\chi_i(S) \\ &= \binom{K}{s}^{-1} \sum_{|S|=s,j\notin S} \mathbb{E}|\mathbf{w}[S]^T\mathbf{z}_i[S]|^m.\end{aligned}$$

Notice that we have used the facts that the events $\chi_i(S)$'s are mutually exclusive and that $\mathbf{z}_i[S]$ and $\chi_i(S)$ are independent. Since the random variable $\mathbf{w}[S]^T\mathbf{z}_i[S]$ has distribution $N(0, \|\mathbf{w}[S]\|_2)$, $\mathbb{E}|\mathbf{w}[S]^T\mathbf{z}_i[S]|^m = \|\mathbf{w}[S]\|_2^m \mathbb{E}|Z|^m \leq 2^{-\frac{m}{2}} m!$. Therefore

$$\mathbb{E}|\mathbf{t}_i[j](\mathbf{w})|^m \leq 2^{-\frac{m}{2}} m! \binom{K}{s}^{-1} \sum_{j\notin S, |S|=s} \|\mathbf{w}[S]\|_2^m.$$

Note that by Lemma A.5, $|||\mathbf{w}[-j]|||_s \geq \|\mathbf{w}[-j]\|_2 \geq \|\mathbf{w}[S]\|_2$ for all S such that $j \notin S$. For $\mathbf{w} \in \mathcal{E}_j(s)$, $|||\mathbf{w}|||_s = 1$ and so $\|\mathbf{w}_S\|_2 \leq 1$, which, further implies that $\|\mathbf{w}[S]\|_2^m \leq \|\mathbf{w}[S]\|_2$. Thus we have

$$\begin{aligned}\mathbb{E}|\mathbf{t}_i[j](\mathbf{w})|^m &\leq 2^{-\frac{m}{2}} m! \binom{K}{s}^{-1} \sum_{j\notin S, |S|=s} \|\mathbf{w}[S]\|_2 \\ &\leq 2^{-\frac{m}{2}} m! \frac{s(K-s)}{K(K-1)} |||\mathbf{w}[-j]|||_s \\ &= 2^{-\frac{m}{2}} m! \frac{s(K-s)}{K(K-1)}\end{aligned}$$

For a fixed j, define

$$U_i(\mathbf{w}) = \mathbf{t}_i[j](\mathbf{w}) - \sqrt{\frac{2}{\pi}}\frac{s}{K}\frac{K-s}{K-1}.$$

Notice that $\mathbb{E}U_i(\mathbf{w}) = 0$. In addition,

$$\mathbb{E}|U_i(\mathbf{w})|^m \leq 2^m \mathbb{E}|\mathbf{t}_i[j](\mathbf{w})|^m \leq \frac{1}{2} \times 4\frac{s}{K}\frac{K-s}{K-1} \times m! \times (\sqrt{2})^{m-2}.$$

By Bernstein's inequality

$$\mathbb{P}\left\{\frac{1}{N}|\sum_{i=1}^{N}U_i(\mathbf{w})|>\frac{s}{K}\epsilon\right\}\leq 2\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{2(4\frac{K-s}{K-1}+\sqrt{2}\epsilon)}\right)\leq 2\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{10}\right),$$

for $0<\epsilon\leq 1/2$. Now let $\{\mathbf{w}_i\}$ be an δ-cover of $\mathcal{E}_j(s)$. Since $\mathcal{E}_j(s)$ is contained in the unit ball $\{\mathbf{w}\in\mathbb{R}^{K-1}:\|\mathbf{w}\|_2\leq 1\}$, there exists a cover such that $|\{\mathbf{w}_l\}|\leq\left(\frac{2}{\delta}+1\right)^{K-1}$. For any $\mathbf{w},\mathbf{w}'\in\mathcal{E}_j(s)$, we have

$$|U_i(\mathbf{w})-U_i(\mathbf{w}')|\leq\sum_{j\notin S,|S|=s}\left|(\mathbf{w}[S]-\mathbf{w}'[S])^T\mathbf{z}_i[S]\right|\chi_i(S).$$

Let Z be a standard Gaussian variable. We have

$$\begin{aligned}\mathbb{P}\left\{\sum_{|S|=s,j\notin S}\left|\mathbf{w}[S]^T\mathbf{z}_i[S]\right|\chi_i(S)>\epsilon\right\}&=\binom{K-1}{s}^{-1}\sum_{|S|=s,j\notin S}\mathbb{P}\left\{\left|\mathbf{w}[S]^T\mathbf{z}_i[S]\right|>\epsilon\right\}\\&=\binom{K-1}{s}^{-1}\sum_{|S|=s,j\notin S}\mathbb{P}\left\{\|\mathbf{w}[S]\|_2|Z|>\epsilon\right\}\\&\leq\mathbb{P}\left\{\|\mathbf{w}\|_2|Z|>\epsilon\right\}.\end{aligned}$$

Let Z_i, $i=1,...,N$, be $i.i.d$ standard Gaussian variables. By the one-sided Bernstein's inequality,

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N}|Z_i|\geq 2\right\}\leq\exp\left(-\frac{N(2-\sqrt{2/\pi})^2}{2(4+\sqrt{2}(2-\sqrt{2/\pi})}\right)\leq\exp\left(-\frac{N}{8}\right).$$

Now let $\delta=\frac{s}{K}\frac{\epsilon}{4}$. Thus

$$\begin{aligned}\mathbb{P}\left\{\sup_{\|\mathbf{w}-\mathbf{w}'\|_2\leq\delta}|\frac{1}{N}\sum_{i=1}^{N}(U_i(\mathbf{w})-U_i(\mathbf{w}'))|>\frac{s}{K}\frac{\epsilon}{2}\right\}&\leq\mathbb{P}\left\{\sup_{\|\mathbf{w}'-\mathbf{w}\|_2\leq\delta}\frac{1}{N}\sum_{i=1}^{N}|U_i(\mathbf{w})-U_i(\mathbf{w}')|>\frac{s}{K}\frac{\epsilon}{2}\right\}\\&\leq\mathbb{P}\left\{\sup_{\|\mathbf{w}'-\mathbf{w}\|_2\leq\delta}\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{w}-\mathbf{w}'\|_2|Z_i|>\frac{s}{K}\frac{\epsilon}{2}\right\}\\&\leq\mathbb{P}\left\{\delta\frac{1}{N}\sum_{i=1}^{N}|Z_i|>\frac{s}{K}\frac{\epsilon}{2}\right\}\leq\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N}|Z_i|>2\right\}\\&\leq\exp\left(-\frac{N}{8}\right).\end{aligned}$$

By triangle inequality

$$\sup_{\|\mathbf{w}'-\mathbf{w}\|_2\leq\delta}|\frac{1}{N}\sum_{i=1}^{N}U_i(\mathbf{w}')|\leq\sup_{\|\mathbf{w}'-\mathbf{w}\|_2\leq\delta}|\frac{1}{N}\sum_{i=1}^{N}(U_i(\mathbf{w})-U_i(\mathbf{w}'))|+|\frac{1}{N}\sum_{i=1}^{N}U_i(\mathbf{w})|.$$

Using a union bound, we have

$$\begin{aligned}\mathbb{P}\{\sup_{\|\mathbf{w}'-\mathbf{w}\|_2\le\delta}\left|\frac{1}{N}\sum_{i=1}^{N}U_i(\mathbf{w}')\right|>\frac{s}{K}\epsilon\} &\le \mathbb{P}\left\{\sup_{\|\mathbf{w}-\mathbf{w}'\|_2\le\delta}\left|\frac{1}{N}\sum_{i=1}^{N}(U_i(\mathbf{w})-U_i(\mathbf{w}'))\right|>\frac{s}{K}\frac{\epsilon}{2}\right\}\\ &\quad+\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}U_i(\mathbf{w})\right|>\frac{s}{K}\frac{\epsilon}{2}\right\}\\ &\le \exp\left(-\frac{N}{8}\right)+2\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{40}\right)\\ &\le 3\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{40}\right),\end{aligned}$$

for $0<\epsilon\le 1$. Now apply union bound again,

$$\begin{aligned}\mathbb{P}\left\{\sup_{\mathbf{w}\in\mathcal{E}_j(s)}\frac{1}{N}|\sum_{i=1}^{N}U_i(\mathbf{w})|>\frac{s}{K}\epsilon\right\} &\le \mathbb{P}\left\{\max_{l}\sup_{||\mathbf{w}-\mathbf{w}_l||_2\le\delta}\frac{1}{N}|\sum_{i=1}^{N}U_i(\mathbf{w})|>\frac{s}{K}\epsilon\right\}\\ &\le 3\left(\frac{8K}{\epsilon s}+1\right)^K\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{40}\right).\end{aligned}$$

(2) For $\mathbf{w}\in\mathcal{F}_j(p)$, under the Bernoulli-Gaussian model:

$$\begin{aligned}\mathbb{E}\,|\mathbf{t}_i[j](\mathbf{w})|^m &= \mathbb{E}|Z|^m\sum_{k=1}^{K-1}\sum_{|S|=k,j\notin S}\|\mathbf{w}[S]\|_2^m\times p^k(1-p)^{K-k}\\ &\le \mathbb{E}|Z|^m p\sum_{k=1}^{K-1}\sum_{|S|=k,j\notin S}\|\mathbf{w}[S]\|_2\times p^{k-1}(1-p)^{K-k}\\ &= \mathbb{E}|Z|^m p(1-p)\sum_{k=0}^{K-2}\sum_{|S|=k+1,j\notin S}\|\mathbf{w}[S]\|_2\times p^k(1-p)^{K-2-k}\\ &= \mathbb{E}|Z|^m p(1-p)\left\|\|\mathbf{w}[-j]\|\right\|_p=\mathbb{E}|Z|^m p(1-p)\\ &\le 2^{-m/2}m!p(1-p).\end{aligned}$$

Notice that we have used the fact that $\|\mathbf{w}[S]\|_2\le\|\mathbf{w}[-j]\|_2\le\left\|\|\mathbf{w}[-j]\|\right\|_p=1$ for all S such that $j\notin S$. For each fixed $\mathbf{w}$, define

$$V_i(\mathbf{w})=\mathbf{t}_i[j](\mathbf{w})-\sqrt{\frac{2}{\pi}}(1-p)p.$$

Now we have

$$\mathbb{E}|V_i(\mathbf{w})|^m\le 2^m\mathbb{E}|\mathbf{t}_i[j](\mathbf{w})|^m\le\frac{1}{2}\times 4p(1-p)\times m!\times(\sqrt{2})^{m-2}.$$

The remaining parts of the proof can be preceeded exactly as in the case of the s-sparse model, noticing that we only need to replace $\frac{s}{K}$ by p, and $\frac{K-s}{K-1}$ by $1-p$.

□

A.4 Dual analysis of $|||.|||_s$ and $|||.|||_p$

In this section, we will characterize the dual norms $|||.|||_s^*$ and $|||.|||_p^*$ by second order cone programs (SOCP). The characterization is helpful for deriving bounds for these special norms in the next section.

Lemma A.4. *For $i \in [\![M]\!]$, let $\mathbf{A}_i$ be an $k_i \times K$ with rank k_i. For $\mathbf{z} \in \mathbb{R}^K$, define*

$$\|\mathbf{z}\|_{\mathbf{A}} = \sum_{i=1}^{M} \|\mathbf{A}_i \mathbf{z}\|_2.$$

Then the dual norm of $\|.\|_{\mathbf{A}}$ is

$$\|\mathbf{v}\|_{\mathbf{A}}^* = \inf \left\{ \max_i \|\mathbf{y}_i\|_2, \mathbf{y}_i \in \mathbb{R}^{k_i}, \sum_{i=1}^{M} \mathbf{A}_i^T \mathbf{y}_i = \mathbf{v} \right\}.$$

Proof.

$$\|\mathbf{v}\|_{\mathbf{A}}^* = \sup_{\mathbf{z} \neq 0} \frac{\mathbf{v}^T \mathbf{z}}{\|\mathbf{z}\|_{\mathbf{A}}} = \sup \left\{ \mathbf{v}^T \mathbf{z} : \|\mathbf{z}\|_{\mathbf{A}} \leq 1 \right\}.$$

Introducing Lagrange multiplier $\lambda \geq 0$ for the inequality constraint, the above problem is equivalent to the following

$$\begin{aligned} \|\mathbf{v}\|_{\mathbf{A}}^* &= \sup_{\mathbf{z}} \left\{ \inf_{\lambda \geq 0} \left\{ \mathbf{v}^T \mathbf{z} + \lambda (1 - \|\mathbf{z}\|_{\mathbf{A}}) \right\} \right\} \\ &= \sup_{\mathbf{z}} \left\{ \inf_{\lambda \geq 0} \left\{ \mathbf{v}^T \mathbf{z} + \lambda (1 - \sum_{i=1}^{M} \|\mathbf{A}_i \mathbf{z}\|_2) \right\} \right\}. \end{aligned}$$

The dual problem is

$$d = \inf_{\lambda \geq 0} \left\{ \sup_{\mathbf{z}} \left\{ \mathbf{v}^T \mathbf{z} + \lambda (1 - \sum_{i=1}^{M} \|\mathbf{A}_i \mathbf{z}\|_2) \right\} \right\}.$$

Notice that $\|\mathbf{A}_i \mathbf{z}\|_2 = \sup\{\mathbf{z}^T \mathbf{A}_i^T \mathbf{u}_i : \|\mathbf{u}_i\|_2 \leq 1\}$. Hence

$$d = \inf_{\lambda \geq 0} \left\{ \lambda + \sup_{\mathbf{z}, \mathbf{u}} \left\{ \mathbf{z}^T (\mathbf{v} - \lambda \sum_{i=1}^{M} \mathbf{A}_i^T \mathbf{u}_i) : \|\mathbf{u}_i\|_2 \leq 1 \right\} \right\}.$$

Since the vector $\mathbf{z}$ can be arbitrary, in order to have a finite value, we must have $\lambda \sum_{i=1}^{M} \mathbf{A}_i^T \mathbf{u}_i = \mathbf{v}$. Now let $\mathbf{y}_i = \lambda \mathbf{u}_i$, the problem becomes

$$d = \inf_{\lambda \geq 0} \left\{ \lambda : \sum_{i=1}^{M} \mathbf{A}_i^T \mathbf{y}_i = \mathbf{v}, \|\mathbf{y}_i\|_2 \leq \lambda \right\}.$$

The above problem is exactly equivalent to

$$\inf \left\{ \max_i \|\mathbf{y}_i\|_2, \mathbf{y}_i \in \mathbb{R}^{k_i}, \sum_{i=1}^{M} \mathbf{A}_i^T \mathbf{y}_i = \mathbf{v} \right\}.$$

Finally, notice that the original problem is convex and strictly feasible. Thus Slater's condition holds and the duality gap is zero. Hence

$$||\mathbf{v}||_{\mathbf{A}}^* = \inf \left\{ \max_i \|\mathbf{y}_i\|_2, \mathbf{y}_i \in \mathbb{R}^{k_i}, \sum_{i=1}^{M} \mathbf{A}_i^T \mathbf{y}_i = \mathbf{v} \right\}.$$

□

The following corollary gives an alternative characterization of $|||.|||_s$ and $|||.|||_p$:

Corollary A.1. *Denote by $\mathbf{y}_S \in \mathbb{R}^{|S|}$ a variable vector indexed by the set S (as opposed to being a subvector of $\mathbf{y}$). For $\mathbf{z} \in \mathbb{R}^m$, we have*

$$|||\mathbf{z}|||_s^* = \inf \left\{ \max_{|S|=s} \|\mathbf{y}_S\|_2 : \mathbf{y}_S \in \mathbb{R}^s, \sum_{|S|=s} \mathbf{E}_S^T \mathbf{y}_S = \mathbf{z} \right\},$$

and

$$|||\mathbf{z}|||_p^* = \inf \left\{ \max_S \|\mathbf{y}_S\|_2 : \mathbf{y}_S \in \mathbb{R}^{|S|}, \sum_{k=0}^{m-1} \texttt{pbinom}(k; m-1, p) \sum_{|S|=k+1} \mathbf{E}_S^T \mathbf{y}_S = \mathbf{z} \right\},$$

where $\mathbf{E}_S = \mathbf{I}[S,]/\binom{m-1}{|S|-1}$ and $\mathbf{I} \in \mathbb{R}^{m \times m}$ is the identity matrix.

Proof. This is simply a direct application of Lemma A.4. □

Corollary A.2. *The dual norms $|||.|||_s^*$ and $|||.|||_p^*$ can be computed via a Second Order Cone Program (SOCP).*

Proof. Introducing additional variable $t \geq 0$, the problem of computing $|||\mathbf{z}|||_s^*$ is equivalent to the following formulation

$$\begin{aligned} \inf_{t, \mathbf{y}_S} \quad & t \;\; s.t. \; \|\mathbf{y}_S\|_2 \leq t \text{ for all } S \text{ such that } |S| = s \\ & \text{and } \sum_{|S|=s} \mathbf{E}_S^T \mathbf{y}_S = \mathbf{z}. \end{aligned}$$

Notice that the above program is already in the standard form of SOCP. The case of $|||.|||_p^*$ can be handled in a similar manner. □

A.5 Inequalities of $\|\|\|.\|\|\|_s$ and $\|\|\|.\|\|\|_p$ and their duals

As demonstrated in the last section, it is in general expensive to compute $\|\|\|.\|\|\|_s^*$ and $\|\|\|.\|\|\|_p^*$. In this section, we will derive sharp and easy-to-compute lower and upper bounds to approximate these quantities.

Lemma A.5. *(Monotonicity of $\|\|\mathbf{z}\|\|_s$ and $\|\|\mathbf{z}\|\|_p$) Let $\mathbf{z} \in \mathbb{R}^m$. $\|\|\mathbf{z}\|\|_1 = \|\mathbf{z}\|_1$ and $\|\|\mathbf{z}\|\|_m = \|\mathbf{z}\|_2$. For $1 \leq l < k \leq m$, we have $\|\|\mathbf{z}\|\|_l \geq \|\|\mathbf{z}\|\|_k$; similarly for $0 < p < q < 1$, $\|\|\mathbf{z}\|\|_p \geq \|\|\mathbf{z}\|\|_q$. Furthermore, the equalities hold iff the vector $\mathbf{z}$ contains at most one non-zero entry.*

Proof. By definition, we have

$$\|\|\mathbf{w}\|\|_1 = \frac{\sum_{|S|=1} \|\mathbf{w}[S]\|_2}{\binom{m-1}{1-1}} = \|\mathbf{w}\|_1.$$

Similarly,

$$\|\|\mathbf{w}\|\|_m = \frac{\sum_{|S|=m} \|\mathbf{w}[S]\|_2}{\binom{m-1}{m-1}} = \|\mathbf{w}\|_2.$$

For $1 \leq k \leq m-1$, let S' be a subset of $[\![m]\!]$ such that $|S'| = k+1$. By triangle inequality

$$\sum_{|S|=k, S \subset S'} \|\mathbf{z}[S]\|_2 \geq k \|\mathbf{z}[S']\|_2,$$

where the equality holds iff $\|\mathbf{z}[S']\|_0 \leq 1$. Thus

$$\sum_{|S'|=k+1} \sum_{|S|=k, S \subset S'} \|\mathbf{z}[S]\|_2 \geq k \sum_{|S'|=k+1} \|\mathbf{z}[S']\|_2,$$

and the equality holds iff $\|\mathbf{z}\|_0 \leq 1$. Notice that the LHS of the above inequality is simply $(m-k)\sum_{|S|=k} \|\mathbf{z}[S]\|_2$. Therefore

$$\|\|\mathbf{z}\|\|_k = \binom{m-1}{k-1}^{-1} \sum_{|S|=k} \|\mathbf{z}[S]\|_2 \geq \binom{m-1}{k}^{-1} \sum_{|S|=k+1} \|\mathbf{z}[S]\|_2 = \|\|\mathbf{z}\|\|_{k+1},$$

and so the inequality holds.

For $\|\|.\|\|_p$, let Y be a random variable that follows the binomial distribution with parameters $m-1$ and p. Observe that $\|\|\mathbf{z}\|\|_p = \mathbb{E}\|\|\mathbf{z}\|\|_{Y+1}$, where the expectation is taken with respect to Y. If $\|\mathbf{z}\|_0 > 1$, $\|\|\mathbf{z}\|\|_k$ is strictly decreasing in k by the first part. Hence, $\|\|\mathbf{z}\|\|_p$ as a function of p is also strictly decreasing on $(0,1)$. Indeed, it can be shown that

$$\frac{d}{dp}\|\|\mathbf{z}\|\|_p = \sum_{k=0}^{m-1} \texttt{pbinom}(k; K-1, p) \left(\|\|\mathbf{z}\|\|_{k+1} - \|\|\mathbf{z}\|\|_k\right) < 0.$$

If $\|\mathbf{z}\|_0 \le 1$, then $|||\mathbf{z}|||_1 = |||\mathbf{z}|||_m$ and so $\frac{d}{dp}|||\mathbf{z}|||_p = 0$. Therefore $|||\mathbf{z}|||_p = |||\mathbf{z}|||_1$ is a constant in p. On the other hand, if $|||\mathbf{z}|||_p = |||\mathbf{z}|||_q$ for $0 < p < q < 1$, by the fact that $\frac{d}{dp}|||\mathbf{z}|||_p \le 0$, we must have $\frac{d}{dp}|||\mathbf{z}|||_p = 0$ and so $|||\mathbf{z}|||_{k-1} = |||\mathbf{z}|||_k$ for all $k \in [\![m]\!]$. Thus $\|\mathbf{z}\|_0 \le 1$.

□

Corollary A.3. *(Monotonicity of $|||\mathbf{z}|||_s^*$ and $|||\mathbf{z}|||_p^*$) Let $\mathbf{z} \in \mathbb{R}^m$. $|||\mathbf{z}|||_1^* = \|\mathbf{z}\|_\infty$ and $|||\mathbf{z}|||_m^* = \|\mathbf{z}\|_2$. For $1 \le i < j \le m$, we have $|||\mathbf{z}|||_i^* \le |||\mathbf{z}|||_j^*$; similarly for $0 < p < q < 1$, $|||\mathbf{z}|||_p^* \le |||\mathbf{z}|||_q^*$. Furthermore, the equalities hold iff the vector $\mathbf{z}$ contains at most one non-zero entry.*

Proof. This is a direct consequence of Lemma A.5 and the dual norm definition $|||\mathbf{z}|||^* = \sup_{\mathbf{y}\neq 0} \frac{\mathbf{z}^T\mathbf{y}}{|||\mathbf{y}|||}$. □

Lemma A.6. *Let $p \in (0,1)$ and $k = \lceil (m-1)p + 1 \rceil$. For any $\mathbf{z} \in \mathbb{R}^m$, we have*

1. $|||\mathbf{z}|||_p \ge |||\mathbf{z}|||_k$.
2. $|||\mathbf{z}|||_p^* \le |||\mathbf{z}|||_k^*$.

Proof. Define the function f with domain on $[1, m]$ as follows: let $f(1) = |||\mathbf{z}|||_1 = \|\mathbf{z}\|_1$; for $i \in [\![m-1]\!]$ and $a \in (i, i+1]$, define

$$f(a) = |||\mathbf{z}|||_i + (|||\mathbf{z}|||_{i+1} - |||\mathbf{z}|||_i)(a - i).$$

It is clear that f is piecewise linear by construction. In addition, by Lemma A.10, f is also convex. Notice that $|||\mathbf{z}|||_p = \mathbb{E}|||\mathbf{z}|||_{Y+1} = \mathbb{E}f(Y+1)$, where Y is a random variable from the binomial distribution with parameters $m-1$ and p. By Jensen's inequality,

$$\mathbb{E}f(Y+1) \ge f(\mathbb{E}Y + 1) = f((m-1)p + 1).$$

Thus by Lemma A.5, $|||\mathbf{z}|||_p \ge |||\mathbf{z}|||_k$ for all $k \ge (m-1)p + 1$. So the first part follows.

To upperbound $|||\mathbf{z}|||_p^*$, notice that if $k \ge (m-1)p + 1$,

$$|||\mathbf{z}|||_p^* = \sup_{\mathbf{w}\neq 0} \frac{\mathbf{w}^T\mathbf{z}}{|||\mathbf{w}|||_p} \le \sup_{\mathbf{w}\neq 0} \frac{\mathbf{w}^T\mathbf{z}}{|||\mathbf{w}|||_k} = |||\mathbf{z}|||_k^*.$$

□

For the following lemmas, the quantities $\tau_m(d, a)$ and $L_m(d, k)$ are defined as in Definition 8.1.

Lemma A.7. *(Approximating $\tau_m(d, a)$) For $d \in [\![m]\!]$ and $a \in (0, m]$:*

$$\tau_m(d, a) \le \sqrt{\frac{da}{m}}.$$

Proof. For $k \in [\![m]\!]$, by Jensen's inequality,

$$\mathbb{E}\sqrt{L_m(d,k)} \leq \sqrt{\mathbb{E}L_m(d,k)} = \sqrt{\frac{dk}{m}}.$$

Note that the last equality follows from the expectation of a hypergeometric random variable. Now suppose $a \in (k-1, k]$. By the above inequality and apply Jensen's inequality one more time, we have

$$\begin{aligned}\tau_m(d,a) &= (k-a)\mathbb{E}\sqrt{L_m(d,k-1)} + (1-(k-a))\mathbb{E}\sqrt{L_m(d,k)} \\ &\leq (k-a)\sqrt{\frac{d(k-1)}{m}} + (1-(k-a))\sqrt{\frac{dk}{m}} = \sqrt{\frac{da}{m}}.\end{aligned}$$

□

Lemma A.8. *(Lower bounds for $|||\mathbf{z}|||_s^*$ and $|||\mathbf{z}|||_p^*$) Let $\mathbf{z} \in \mathbb{R}^m$. We have*

1. *For $s \in [\![m]\!]$,*

$$|||\mathbf{z}|||_s^* \geq \frac{s}{m} \max_{T \subset [\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\tau_m(|T|, s)} \geq \max\left(\|\mathbf{z}\|_\infty, \sqrt{\frac{s}{m}} \max_{T \subset [\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\sqrt{|T|}}\right).$$

2. *For $p \in (0,1)$,*

$$\begin{aligned}|||\mathbf{z}|||_p^* &\geq p \max_{T \subset [\![m]\!]} \left\{ \Big(\sum_{k=0}^{m} \texttt{pbinom}(k, m, p) \tau_m(|T|, k) \Big)^{-1} \|\mathbf{z}[T]\|_1 \right\} \\ &\geq p \max_{T \subset [\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\tau_m(|T|, pm)} = \max\left(\|\mathbf{z}\|_\infty, \sqrt{p} \max_{T \subset [\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\sqrt{|T|}}\right).\end{aligned}$$

Proof. (1) Note that by definition,

$$|||\mathbf{z}|||_s^* = \sup_{\mathbf{w}} \frac{\mathbf{z}^T\mathbf{w}}{|||\mathbf{w}|||_s}$$

Let $d \in [\![m]\!]$ and $T \subset [\![m]\!]$ such that $|T| = d$. Define $\mathbf{w} \in \mathbb{R}^m$ such that $\mathbf{w}[i] = 1$ for $i \in T$

and $\mathbf{w}[i] = 0$ for $i \in T^c$. We have:

$$\begin{aligned}
|||\mathbf{w}|||_s = \binom{m-1}{s-1}^{-1} \sum_{|S|=s} \|\mathbf{w}[S]\|_2 &= \binom{m-1}{s-1}^{-1} \sum_{l=\max(0,s+d-m)}^{\min(s,d)} \sum_{|S|=s,|S\cap T|=l} \|\mathbf{w}[S]\|_2 \\
&= \binom{m-1}{s-1}^{-1} \sum_{l=\max(0,s+d-m)}^{\min(s,d)} \sum_{|S|=s,|S\cap T|=l} \sqrt{l} \\
&= \binom{m-1}{s-1}^{-1} \sum_{l=\max(0,s+d-m)}^{\min(s,d)} \binom{d}{l}\binom{m-d}{s-l}\sqrt{l} \\
&= \frac{m}{s}\mathbb{E}\sqrt{L_m(d,s)} = \frac{m}{s}\tau_m(d,s).
\end{aligned}$$

Thus for all $d \in [\![m]\!]$ and any subset T such that $|T| = d$, we have shown

$$|||\mathbf{z}|||_s^* \geq \frac{s}{m}\frac{\|\mathbf{z}[T]\|_1}{\tau_m(d,s)}.$$

Note that if $d = 1$, $\mathbb{E}\sqrt{L_m(d,s)} = \frac{s}{m}$. Therefore

$$\frac{s}{m}\frac{\|\mathbf{z}[T]\|_1}{\tau_m(d,s)} \geq \|\mathbf{z}\|_\infty,$$

Moreover, by Lemma A.7,

$$\tau_m(d,s) \leq \sqrt{\frac{ds}{m}}.$$

Hence we have

$$\frac{s}{m}\frac{\|\mathbf{z}[T]\|_1}{\tau_m(d,s)} \geq \sqrt{\frac{s}{m}}\frac{\|\mathbf{z}[T]\|_1}{\sqrt{d}},$$

and the first part of the claim follows.

(2) For the same $\mathbf{w} \in \mathbb{R}^m$ defined previously,

$$\begin{aligned}
|||\mathbf{w}|||_p &= \sum_{k=0}^{m-1} \texttt{pbinom}(k, m-1, p) |||\mathbf{w}|||_{k+1} \\
&= m \sum_{k=0}^{m-1} \texttt{pbinom}(k, m-1, p) \frac{\tau_m(d, k+1)}{k+1} \\
&= m \sum_{k=0}^{m-1} \binom{m-1}{k} p^k (1-p)^{m-k-1} \frac{1}{k+1} \tau_m(d, k+1) \\
&= \frac{1}{p} \sum_{k=0}^{m-1} \binom{m}{k+1} p^{k+1} (1-p)^{m-(k+1)} \tau_m(d, k+1) \\
&= \frac{1}{p} \sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} \tau_m(d, k).
\end{aligned}$$

Thus for all $d \in [\![m]\!]$ and any subset T such that $|T| = d$, we have shown

$$|||\mathbf{z}|||_p^* \geq p \Big(\sum_{k=0}^{m} \texttt{pbinom}(k, m, p) \tau_m(d, k) \Big)^{-1} \|\mathbf{z}[T]\|_1.$$

Next, we will show

$$\sum_{k=0}^{m} \texttt{pbinom}(k, m, p) \tau_m(d, k) \leq \tau_m(d, pm).$$

To this end, let us first notice that the LHS quantity is a binomial average of $\tau_m(d, k)$ with respect to k. By construction, $\tau_m(d, .)$ is piecewise linear. Furthermore, $\tau_m(d, .)$ is also concave by Lemma A.11. Now let Y be a random variable having the binomial distribution with parameters m and p. By Jensen's inequality,

$$\sum_{k=0}^{m} \texttt{pbinom}(k, m, p) \tau_m(d, k) = \mathbb{E} \tau_m(d, Y) \leq \tau_m(d, \mathbb{E}Y) = \tau_m(d, mp).$$

In particular, if $d = 1$, it is easy to see that $\tau_m(d, mp) = p$. So

$$p \left(\max_{T \subset [\![m]\!], |T|=1} \Big(\sum_{k=0}^{m} \texttt{pbinom}(k, m, p) \tau_m(|T|, k) \Big)^{-1} \|\mathbf{z}[T]\|_1 \right) \geq \|\mathbf{z}\|_\infty.$$

On the other hand, by Lemma A.7,

$$\tau_m(d, pm) \leq \sqrt{\frac{d}{m}} \sqrt{pm} = \sqrt{pd}.$$

Therefore

$$p\Big(\sum_{k=0}^{m} \texttt{pbinom}(k,m,p)\tau_m(d,k)\Big)^{-1}\|\mathbf{z}[T]\|_1 \geq \sqrt{p}\frac{\|\mathbf{z}[T]\|_1}{\sqrt{d}},$$

and the proof is complete. □

Lemma A.9. *(Upper bounds for $|||\mathbf{z}|||_s^*$ and $|||\mathbf{z}|||_p^*$) Let $\mathbf{z} \in \mathbb{R}^m$.*

1. *For $s \in [\![m]\!]$,*
$$|||\mathbf{z}|||_s^* \leq \max_{|S|=s} \|\mathbf{z}[S]\|_2.$$
2. *For $p \in (0,1)$,*
$$|||\mathbf{z}|||_p^* \leq \max_{|S|=k} \|\mathbf{z}[S]\|_2,$$
where $k = \lceil p(m-1)+1 \rceil$.

Proof. To establish the upper bound, we will use the equivalent formulation of $|||.|||_s^*$ in Corollary A.1. For $S \subset [\![m]\!]$ of size s, as in Corollary A.1, let $\mathbf{E}_S = \mathbf{I}[S,]/\binom{m-1}{s-1}$ where $\mathbf{I} \in \mathbb{R}^{m\times m}$ is the identity matrix. If we set $\mathbf{y}_S = \mathbf{z}[S]$, then $\sum_{|S|=s} \mathbf{E}_S^T \mathbf{y}_S = \mathbf{z}$ and so $\{\mathbf{y}_S\}$ is feasible. Therefore

$$|||\mathbf{z}|||_s^* \leq \max_{|S|=s} \|\mathbf{z}[S]\|_2.$$

The upperbound of $|||\mathbf{z}|||_p^*$ follows from the inequality $|||\mathbf{z}|||_p^* \leq |||\mathbf{z}|||_k^*$ for $k = \lceil p(m-1)+1 \rceil$ by the second part of Lemma A.6. □

Corollary A.4. *(1-sparse vectors) Let $\mathbf{z} = (z, 0, ..., 0)^T \in \mathbb{R}^m$. We have*

$$|||\mathbf{z}|||_s^* = |||\mathbf{z}|||_p^* = |z|.$$

Proof. These are direct consequences of Lemma A.8 and Lemma A.9. □

Corollary A.5. *(All-constant vectors) Let $\mathbf{z} \in \mathbb{R}^m$ be such that $\mathbf{z}[i] = z$ for all $i \in [\![m]\!]$. We have*

1. $|||\mathbf{z}|||_s^* = \sqrt{s}|z|$.
2. $|||\mathbf{z}|||_p^* = mp\Big(\sum_{k=0}^{m} \texttt{pbinom}(k,m,p)\sqrt{k}\Big)^{-1}|z|$.

Proof. First of all, note that $L(m,k) = k$ and $\mathbb{E}\sqrt{L(m,k)} = \sqrt{k}$. Thus by Lemma A.8 and A.9, we have

$$|||\mathbf{z}|||_s^* = \sqrt{s}|z|.$$

So the first part of the claim is verified. Next, by Lemma A.8,

$$|||\mathbf{z}|||_p^* \geq mp\Big(\sum_{k=0}^{m} \texttt{pbinom}(k,m,p)\sqrt{k}\Big)^{-1}|z|.$$

On the other hand, for S such that $|S| = s$, we can define

$$\mathbf{y}_S = \frac{mp}{\sqrt{s}}\Big(\sum_{k=0}^{m} \texttt{pbinom}(k, m, p)\sqrt{k}\Big)^{-1}(z, ..., z)^T \in \mathbb{R}^s,$$

For notation simplicity, let $c = \frac{1}{mp}\Big(\sum_{k=0}^{m} \texttt{pbinom}(k, m, p)\sqrt{k}\Big)$. As in Corollary A.1, for $S \subset [\![m]\!]$, let $\mathbf{E}_S = \mathbf{I}[S,]/\binom{m-1}{|S|-1}$. For $i \in [\![m]\!]$, we have

$$\begin{aligned}\sum_{k=0}^{m-1} \texttt{pbinom}(k; m-1, p) \sum_{|S|=k+1} (\mathbf{E}_S^T \mathbf{y}_S)[i] &= c^{-1} \sum_{k=0}^{m-1} \texttt{pbinom}(k; m-1, p) \frac{1}{\sqrt{k+1}} \\ &= c^{-1} \frac{z}{mp} \sum_{k=0}^{m} \texttt{pbinom}(k; m, p)\sqrt{k} = z.\end{aligned}$$

Thus by Corollary A.1,

$$|||\mathbf{z}|||_p^* \le \max_S \|\mathbf{y}_S\|_2 = mp\Big(\sum_{k=0}^{m} \texttt{pbinom}(k, m, p)\sqrt{k}\Big)^{-1}|z|,$$

and the proof is complete. □

Lemma A.10. *(Convexity of $|||\mathbf{z}|||_k$) Let $\mathbf{z} \in \mathbb{R}^m$, where $m \ge 3$. For $k \in [\![m-2]\!]$, we have the following inequality*

$$|||\mathbf{z}|||_k + |||\mathbf{z}|||_{k+2} \ge 2|||\mathbf{z}|||_{k+1}. \tag{A.20}$$

Proof. We will first show that the claim is true for $k = m-2$. Notice that in this case $|||\mathbf{z}|||_{k+2} = |||\mathbf{z}|||_m = \|\mathbf{z}\|_2$. If $\|\mathbf{z}\|_2 = 0$, the inequality (A.20) is trivially true. Now suppose $\|\mathbf{z}\|_2 > 0$, dividing both sides of the inequality by $\|\mathbf{z}\|_2$, we have

$$\binom{m-1}{m-3}^{-1} \sum_{|S|=m-2} \frac{\|\mathbf{z}[S]\|_2}{\|\mathbf{z}\|_2} + 1 \ge 2\binom{m-1}{m-2}^{-1} \sum_{|S|=m-1} \frac{\|\mathbf{z}[S]\|_2}{\|\mathbf{z}\|_2}.$$

Now let $\mathbf{x} = (x_1, ..., x_m)^T \in \mathbb{R}^m$ be such that $x_i = \mathbf{z}[i]^2/\|\mathbf{z}\|_2^2$. It suffices to show

$$\sum_{|S|=m-2} \left(\sum_{i\in S} x_i\right)^{1/2} + \frac{(m-1)(m-2)}{2} \ge (m-2)\sum_{i=1}^{m} \sqrt{1-x_i}, \tag{A.21}$$

for all $\mathbf{x} \ge 0$ entry-wise such that $\sum_i x_i = 1$. We will now prove the above inequality by induction on m. First of all, notice that for the base case where $m = 3$, we need to show:

$$\sqrt{x_1} + \sqrt{x_2} + \sqrt{x_3} + 1 \ge \sqrt{1-x_1} + \sqrt{1-x_2} + \sqrt{1-x_3},$$

with the constraints $x_i \geq 0$ and $x_1 + x_2 + x_3 = 1$. For fixed x_3, let

$$f(x_1) = \sqrt{x_1} + \sqrt{1 - x_1 - x_3} + \sqrt{x_3} + 1 - \sqrt{x_1 + x_3} - \sqrt{1 - x_1} - \sqrt{1 - x_3}.$$

We will show that $f(x_1)$ is minimized at $x_1 = 0$ or $x_1 = 1 - x_3$. Suppose now $x_1 > 0$. Taking derivative with respect to x_1:

$$f'(x_1) = \frac{1}{2}\left(\frac{1}{\sqrt{x_1}} - \frac{1}{\sqrt{1 - x_1 - x_3}} - \frac{1}{\sqrt{x_1 + x_3}} + \frac{1}{\sqrt{1 - x_1}}\right).$$

Let $l(x_1) = \frac{1}{\sqrt{x_1}} - \frac{1}{\sqrt{x_1 + x_3}}$. Note that $f'(x_1) = \frac{1}{2}l(x_1) - \frac{1}{2}l(1 - x_3 - x_1)$. Now we have

$$l'(x_1) = \frac{1}{2}(x_1 + x_3)^{-3/2} - \frac{1}{2}{x_1}^{-3/2}.$$

So $l(x_1)$ is decreasing on $(0, 1 - x_3)$ and by symmetry the function $l(1 - x_3 - x_1)$ is increasing on $(0, 1 - x_3)$. On the other hand, since $\lim_{x_1 \downarrow 0^+} l(x_1) = +\infty$ and $\lim_{x_1 \downarrow 0^+} l(1 - x_3 - x_1) = -\infty$, we know that $f'(x_1) > 0$ on $(0, \frac{1-x_3}{2})$ and < 0 on $(\frac{1-x_3}{2}, 1 - x_3)$. Thus, the minimum of f can only be attained at the boundaries, i.e. $x_1 = 0$ or $x_1 = 1 - x_3$. In either case we have

$$\begin{aligned}
&\sqrt{x_1} + \sqrt{x_2} + \sqrt{x_3} + 1 - \sqrt{1 - x_1} - \sqrt{1 - x_2} - \sqrt{1 - x_3} \\
\geq &\sqrt{x_2} + \sqrt{x_3} - \sqrt{1 - x_2} - \sqrt{1 - x_3} = 0,
\end{aligned}$$

as $x_2 + x_3 = 1$. So we establish (A.21) for $m = 3$.

Suppose (A.21) is also true for $m = n - 1$. For $m = n$, similar to the $m = 3$ case, for fixed $x_3, ..., x_n$, define

$$f(x_1) = \sum_{|S|=n-2} \left(\sum_{i \in S} x_i\right)^{1/2} + \frac{(n-1)(n-2)}{2} - (n-2)\sum_{i=1}^{n} \sqrt{1 - x_i},$$

subject to $x_i \geq 0$ and $\sum_i x_i = 1$. Again, we will show f attains its minimum at either $x_1 = 0$ or $x_1 = 1 - \sum_{i=3}^{n} x_i$. Notice that

$$\begin{aligned}
\sum_{|S|=n-2} \Big(\sum_{j \in S} x_j\Big)^{1/2} = &\sum_{|S|=n-3, 1, 2 \notin S} \Big(x_1 + \sum_{j \in S} x_j\Big)^{1/2} + \sum_{|S|=n-4, 1, 2 \notin S} \Big(x_1 + x_2 + \sum_{j \in S} x_j\Big)^{1/2} \\
&+ \sum_{|S|=n-3, 1, 2 \notin S} \Big(x_2 + \sum_{j \in S} x_j\Big)^{1/2} + \Big(\sum_{j=3}^{n} x_j\Big)^{1/2} \\
= &\sum_{i=3}^{n} \Big(x_1 + \sum_{j=3}^{n} x_j - x_i\Big)^{1/2} + \sum_{3 \leq i < j \leq n} (1 - x_i - x_j)^{1/2} \\
&+ \sum_{i=3}^{n} (1 - x_1 - x_i)^{1/2} + \Big(\sum_{j=3}^{n} x_j\Big)^{1/2}.
\end{aligned}$$

In addition,

$$\sum_{i=1}^{n}(1-x_i)^{1/2} = (1-x_1)^{1/2} + \Big(x_1 + \sum_{j=3}^{n} x_j\Big)^{1/2} + \sum_{i=3}^{n}(1-x_i)^{1/2}.$$

Taking derivative with respect to x_1,

$$\begin{aligned} f'(x_1) =& \frac{1}{2}\Big(\sum_{i=3}^{n}\Big(x_1 + \sum_{j=3}^{n} x_j - x_i\Big)^{-1/2} - \sum_{i=3}^{n}(1-x_1-x_i)^{-1/2} \\ &+ (n-2)(1-x_1)^{-1/2} - (n-2)\Big(x_1 + \sum_{i=3}^{n} x_i\Big)^{-1/2}\Big). \end{aligned}$$

Now let

$$l(x_1) = \sum_{i=3}^{n}\Big(x_1 + \sum_{j=3}^{n} x_j - x_i\Big)^{-1/2} - (n-2)\Big(x_1 + \sum_{j=3}^{n} x_j\Big)^{-1/2}.$$

So $2f'(x_1) = l(x_1) - l(1 - \sum_{i=3}^{n} x_i - x_1)$. Again

$$l'(x_1) = -\frac{1}{2}\sum_{i=3}^{n}\Big(x_1 + \sum_{j=3}^{n} x_j - x_i\Big)^{-3/2} + \frac{n-2}{2}\Big(x_1 + \sum_{j=3}^{n} x_j\Big)^{-3/2}.$$

It is easy to see that $l'(x_1) < 0$ and so $l(x_1)$ is decreasing on $(0, 1 - \sum_{i=3}^{n} x_i - x_1)$. On the other hand $\lim_{x_1 \downarrow 0^+} l(x_1) = +\infty$. By symmetry $f'(x_1) > 0$ on $\big(0, \frac{1}{2}(1 - \sum_{i=3}^{n} x_i - x_1)\big)$ and < 0 on $\big(0, \frac{1}{2}(1 - \sum_{i=3}^{n} x_i - x_1)\big)$. So f attains its minimum at $x_1 = 0$ or $x_1 = 1 - \sum_{i=3}^{n} x_i$. Hence we have

$$\begin{aligned} &\sum_{|S|=n-2}\Big(\sum_{i\in S} x_i\Big)^{1/2} + \frac{(n-1)(n-2)}{2} - (n-2)\sum_{i=1}^{n}(1-x_i)^{1/2} \\ \geq& \Big(\sum_{|S|=n-3, 1\notin S} + \sum_{|S|=n-2, 1\notin S}\Big)\Big(\sum_{j\in S} x_j\Big)^{1/2} + \frac{(n-2)(n-3)}{2} - (n-2)\sum_{i=2}^{n}(1-x_i)^{1/2}. \quad \text{(A.22)} \end{aligned}$$

By the induction assumption that (A.21) holds when $m = n-1$, we have

$$\sum_{|S|=n-3, 1\notin S}\Big(\sum_{j\in S} x_j\Big)^{1/2} + \frac{(n-2)(n-3)}{2} \geq (n-3)\sum_{i=2}^{n}(1-x_i)^{1/2}.$$

Thus (A.22) is greater than or equal to

$$\begin{aligned} &-\frac{(n-2)(n-3)}{2} + \sum_{|S|=n-2, 1\notin S}\Big(\sum_{j\in S} x_j\Big)^{1/2} + \frac{(n-1)(n-2)}{2} - (n-2) - \sum_{i=2}^{n}(1-x_i)^{1/2} \\ =& \sum_{|S|=n-2, 1\notin S}\Big(\sum_{j\in S} x_j\Big)^{1/2} - \sum_{i=2}^{n}(1-x_i)^{1/2} = \sum_{i=2}^{n}(1-x_i)^{1/2} - \sum_{i=2}^{n}(1-x_i)^{1/2} = 0. \end{aligned}$$

Thus we have verified the claim that (A.21) and hence (A.20) holds for $k = m - 2$ for all $m \geq 3$. To establish the case for general $1 \leq k \leq m - 2$, we again perform induction on the (m, k)-tuple. Note that the base case $m = 3$ and $k = 1$ has been previously proved. Suppose (A.20) holds for $m = n - 1$ and $1 \leq k \leq n - 3$. Now consider $m = n$ and $1 \leq k < n - 2$. Notice that

$$\begin{aligned}
|||\mathbf{z}|||_k =& \frac{1}{n-k}\binom{n-1}{k-1}^{-1} \sum_{|T|=n-1} \sum_{|S|=k, S\subset T} \|\mathbf{z}[S]\|_2 \\
=& (n-1)\binom{n-2}{k-1}^{-1} \sum_{|T|=n-1} \sum_{|S|=k, S\subset T} \|\mathbf{z}[S]\|_2 \\
=& (n-1) \sum_{|T|=n-1} |||\mathbf{z}[T]|||_k.
\end{aligned}$$

By the induction assumption, for all T such that $|T| = n - 1$, we have:

$$|||\mathbf{z}[T]|||_k + |||\mathbf{z}[T]|||_{k+2} \geq 2|||\mathbf{z}[T]|||_{k+1}.$$

Therefore

$$|||\mathbf{z}|||_k + |||\mathbf{z}|||_{k+2} - 2|||\mathbf{z}|||_{k+1} = (n-1) \sum_{|T|=n-1} (|||\mathbf{z}[T]|||_k + |||\mathbf{z}[T]|||_{k+2} - 2|||\mathbf{z}[T]|||_{k+1}) \geq 0.$$

Thus the claim also holds for $m = n$ and $1 \leq k < n - 2$, completing the proof. □

A.6 Miscellaneous

Lemma A.11. *(Concavity of $\mathbb{E}\sqrt{L_m(d,k)}$) Let $d \in [\![m]\!]$. For $k \in [\![m-2]\!]$, we have*

$$\mathbb{E}\sqrt{L_m(d,k)} + \mathbb{E}\sqrt{L_m(d,k+2)} \leq 2\mathbb{E}\sqrt{L_m(d,k+1)}. \tag{A.23}$$

where the geometric random variable $L_m(d,k)$ is defined as in Definition 8.1.

Proof. Suppose we are now sampling without replacement from a pool of numbers with d 1's and $m - d$ 0's. For $i \in [\![m]\!]$, denote by $X_i \in \{0, 1\}$ the i-th outcome. It is easy to see that $L_m(d,k)$ and $\sum_{i=1}^{k} X_i$ have the same distribution. To show (A.23), it suffices to prove the following conditional expectation inequality:

$$\sqrt{L_m(d,k)} + \mathbb{E}[\sqrt{L_m(d,k+2)} \mid L_m(d,k)] \leq 2\mathbb{E}[\sqrt{L_m(d,k+1)} \mid L_m(d,k)]$$

Note that the above inequality follows if for all $0 \leq a \leq \min(d,k)$:

$$\sqrt{a} + \mathbb{E}\sqrt{a + X_{k+1} + X_{k+2}} \leq 2\mathbb{E}\sqrt{a + X_{k+1}}$$

It is easy to see that

$$\mathbb{E}\sqrt{a+X_{k+1}} = \frac{d-a}{m-k}\sqrt{a+1} + (1-\frac{d-a}{m-k})\sqrt{a}.$$

$$\begin{aligned}\mathbb{E}\sqrt{a+X_{k+1}+X_{k+2}} = &\frac{d-a}{m-k}\times\frac{d-a-1}{m-k-1}\sqrt{a+2} + 2\times\frac{d-a}{m-k}\times\frac{m-k-(d-a)}{m-k-1}\sqrt{a+1}\\ &+\frac{m-k-(d-a)}{m-k}\times\frac{m-k-(d-a)-1}{m-k-1}\sqrt{a}.\end{aligned}$$

By elementary algebra, it can be shown that

$$\begin{aligned}&2\mathbb{E}\sqrt{a+X_{k+1}} - \sqrt{a} - \mathbb{E}\sqrt{a+X_{k+1}+X_{k+2}}\\ =&\frac{d-a}{m-k}\times\frac{d-a-1}{m-k-1}\times(2\sqrt{a+1}-\sqrt{a+2}-\sqrt{a}) \geq 0,\end{aligned}$$

The inequality follows since $f(x)=\sqrt{x}$ is a concave function. Thus the proof is complete. □

Lemma A.12. *Let $\mathbf{x}(t) = (x_1(t), ..., x_m(t))^T \in \mathbb{R}^m$ be an m-dimensional function on $[0,\epsilon)$ such that: (1) $x_1(0)=1$ and for all $i \geq 2$, $x_i(0)=0$; (2) The derivative $\dot{x}_i(t)$ exists and is bounded for all $t \in (0,\epsilon)$. We have*

$$\lim_{t\downarrow 0^+}\frac{\|\mathbf{x}(t)\|_2 - \|\mathbf{x}(0)\|_2}{t} = \lim_{t\downarrow 0^+}\dot{\mathbf{x}}_1(t).$$

Proof.

$$\begin{aligned}\lim_{t\downarrow 0^+}\frac{\|\mathbf{x}(t)\|_2 - \|\mathbf{x}(0)\|_2}{t} &= \lim_{t\downarrow 0^+}\frac{(\sum_{i=1}^m x_i^2(t))^{1/2}-1}{t}\\ &= \lim_{t\downarrow 0^+}\frac{\sum_{i=1}^m x_i^2(t)-1}{t}((\sum_{i=1}^m x_i^2(t))^{1/2}+1)^{-1}\\ &= \frac{1}{2}\lim_{t\downarrow 0^+}\frac{\sum_{i=1}^m x_i^2(t)-1}{t}\\ &= \frac{1}{2}(\lim_{t\downarrow 0^+}\frac{x_1^2(t)-1}{t} + \sum_{i=2}^m\lim_{t\downarrow 0^+}\frac{x_i^2(t)}{t})\\ &= \frac{1}{2}(\lim_{t\downarrow 0^+}\frac{x_1(t)-1}{t}(x_1(t)+1) + \sum_{i=2}^m\lim_{t\downarrow 0^+}\frac{x_i^2(t)}{t})\\ &= \lim_{t\downarrow 0^+}\frac{x_1(t)-1}{t} + \frac{1}{2}\sum_{i=2}^m\lim_{t\downarrow 0^+}\frac{x_i^2(t)}{t}.\end{aligned}$$

By mean value theorem, for each $t \in (0,\epsilon)$, there exists $\delta_t \in (0,t)$ such that $x_1(t)-1 = \dot{x}_1(\delta_t)t$. Thus the first term simply becomes $\lim_{t\downarrow 0^+}\dot{x}_1(t)$. By the same argument, for each $i \in \{2,...,m\}$, $x_i(t) = \dot{x}_i(\delta_t)t$ for some $\delta_t \in (0,t)$. Since $\dot{x}_i(t)$ is bounded, we have

$$\lim_{t\downarrow 0^+}\frac{x_i^2(t)}{t} = \lim_{t\downarrow 0^+}\dot{x}_i(\delta_t)^2 t = 0.$$

Therefore the claim is verified. □

Lemma A.13. *Let* $\mathbf{x}(t) = (x_1(t), ..., x_m(t))^T \in \mathbb{R}^m$ *be an* m*-dimensional function on* $[0, \epsilon)$ *such that: (1)* $x_i(0) = 0$ *for all* $i = 1, ..., m$*; (2) The derivative* $\dot{x}_i(t)$ *exists for all* $t \in (0, \epsilon)$*. We have*

$$\lim_{t \downarrow 0^+} \frac{\|\mathbf{x}(t)\|_2}{t} = \| \lim_{t \downarrow 0^+} \dot{\mathbf{x}}(t) \|_2.$$

Proof.

$$\lim_{t \downarrow 0^+} \frac{\|\mathbf{x}(t)\|_2}{t} = \lim_{t \downarrow 0^+} (\sum_{i=1}^{m} (\frac{x_i(t)}{t})^2)^{1/2} = (\sum_{i=1}^{m} (\lim_{t \downarrow 0^+} \frac{x_i(t)}{t})^2)^{1/2} = \| \lim_{t \downarrow 0^+} \dot{\mathbf{x}}(t) \|_2.$$

□

Lemma A.14. *Let* $\mathbf{a} = (a_1, ..., a_m)^T \in \mathbb{R}^m$ *where* $a_1 \neq 0$ *and* $\mathbf{x}(t) = (x_1(t), ..., x_m(t))^T \in \mathbb{R}^m$ *be an* m*-dimensional function on* $[0, \epsilon)$ *such that: (1)* $x_1(0) = 1$ *and for all* $i \geq 2$*,* $x_i(0) = 0$*; (2) The derivative* $\dot{x}_i(t)$ *exists and is bounded for all* $t \in (0, \epsilon)$*. We have*

$$\lim_{t \downarrow 0^+} \frac{|\mathbf{a}^T \mathbf{x}(t)| - |a_1|}{t} = |a_1| \lim_{t \downarrow 0^+} \dot{x}_1(t) + \boldsymbol{sgn}(a_1) \sum_{i=2}^{m} a_i \lim_{t \downarrow 0^+} \dot{x}_i(t).$$

Proof. Without loss of generality, assume $a_1 > 0$. Since $x_1(0) = 1$ and for all $i \geq 2$, $x_i(0) = 0$, by continuity, for sufficiently small t, we have

$$\frac{|\mathbf{a}^T \mathbf{x}(t)| - |a_1|}{t} = \frac{|a_1 x_1(t) + \sum_{i=2}^{m} a_i x_i(t)| - a_1}{t} = \frac{a_1 x_1(t) - a_1 + \sum_{i=2}^{m} a_i x_i(t)}{t}.$$

Therefore, by the same argument in the proof of Lemma A.12,

$$\begin{aligned} \lim_{t \downarrow 0^+} \frac{|\mathbf{a}^T \mathbf{x}(t)| - |a_1|}{t} &= \lim_{t \downarrow 0^+} \frac{a_1 x_1(t) - a_1}{t} + \lim_{t \downarrow 0^+} \sum_{i=2}^{m} \frac{a_i x_i(t)}{t} \\ &= a_1 \lim_{t \downarrow 0^+} \dot{x}_1(t) + \sum_{i=2}^{m} a_i \lim_{t \downarrow 0^+} \dot{x}_i(t). \end{aligned}$$

□

Lemma A.15. *Let* $\mathbf{a} = (a_1, ..., a_m)^T \in \mathbb{R}^m$ *and* $\mathbf{x}(t) = (x_1(t), ..., x_m(t))^T \in \mathbb{R}^m$ *be an* m*-dimensional function on* $[0, \epsilon)$ *such that: (1)* $x_i(0) = 0$ *for all* $i = 1, ..., m$*; (2) The derivative* $\dot{x}_i(t)$ *exists for all* $t \in (0, \epsilon)$*. We have*

$$\lim_{t \downarrow 0^+} \frac{|\mathbf{a}^T \mathbf{x}(t)|}{t} = \left| \sum_{i=1}^{m} a_i \lim_{t \downarrow 0^+} \dot{x}_i(t) \right|.$$

Proof. The proof is similar to that of Lemma A.13. □