University Name

Doctoral Thesis

# Thesis Title

*Author:*
John Smith

*Supervisor:*
Dr. James Smith

*A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy*

*in the*

Research Group Name
Department or School Name

January 19, 2023

# Declaration of Authorship

I, John Smith, declare that this thesis titled, "Thesis Title" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"*Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I got a good job in journalism.*"

Dave Barry

UNIVERSITY NAME

# *Abstract*

Faculty Name
Department or School Name

Doctor of Philosophy

**Thesis Title**

by John Smith

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

# *Acknowledgements*

The acknowledgments and the people to thank go here, don't forget to include your project advisor...

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| **LAH** | **L**ist **A**bbreviations **H**ere |
| **WSF** | **W**hat (it) **S**tands **F**or |

# Physical Constants

Speed of Light $\quad c_0 = 2.997\,924\,58 \times 10^8\,\mathrm{m\,s^{-1}}$ (exact)

# List of Symbols

| | | |
|---|---|---|
| $a$ | distance | m |
| $P$ | power | W ($\mathrm{J\,s^{-1}}$) |
| $\omega$ | angular frequency | rad |

*For/Dedicated to/To my...*

# Chapter 1

# Chapter Title Here

## 1.1 Welcome and Thank You

Welcome to this LaTeX Thesis Template, a beautiful and easy to use template for writing a thesis using the LaTeX typesetting system.

If you are writing a thesis (or will be in the future) and its subject is technical or mathematical (though it doesn't have to be), then creating it in LaTeX is highly recommended as a way to make sure you can just get down to the essential writing without having to worry over formatting or wasting time arguing with your word processor.

LaTeX is easily able to professionally typeset documents that run to hundreds or thousands of pages long. With simple mark-up commands, it automatically sets out the table of contents, margins, page headers and footers and keeps the formatting consistent and beautiful. One of its main strengths is the way it can easily typeset mathematics, even *heavy* mathematics. Even if those equations are the most horribly twisted and most difficult mathematical problems that can only be solved on a super-computer, you can at least count on LaTeX to make them look stunning.

## 1.2 Learning LaTeX

LaTeX is not a WYSIWYG (What You See is What You Get) program, unlike word processors such as Microsoft Word or Apple's Pages. Instead, a document written for LaTeX is actually a simple, plain text file that contains *no formatting*. You tell LaTeX how you want the formatting in the finished document by writing in simple commands amongst the text, for example, if I want to use *italic text for emphasis*, I write the `\emph{text}` command and put the text I want in italics in between the curly braces. This means that LaTeX is a "mark-up" language, very much like HTML.

### 1.2.1 A (not so short) Introduction to LaTeX

If you are new to LaTeX, there is a very good eBook – freely available online as a PDF file – called, "The Not So Short Introduction to LaTeX". The book's title is typically shortened to just *lshort*. You can download the latest version (as it is occasionally updated) from here: http://www.ctan.org/tex-archive/info/lshort/english/lshort.pdf

It is also available in several other languages. Find yours from the list on this page: http://www.ctan.org/tex-archive/info/lshort/

It is recommended to take a little time out to learn how to use LaTeX by creating several, small 'test' documents, or having a close look at several templates on: http://www.LaTeXTemplates.com

Making the effort now means you're not stuck learning the system when what you *really* need to be doing is writing your thesis.

### 1.2.2 A Short Math Guide for LaTeX

If you are writing a technical or mathematical thesis, then you may want to read the document by the AMS (American Mathematical Society) called, "A Short Math Guide for LaTeX". It can be found online here: http://www.ams.org/tex/amslatex.html under the "Additional Documentation" section towards the bottom of the page.

### 1.2.3 Common LaTeX Math Symbols

There are a multitude of mathematical symbols available for LaTeX and it would take a great effort to learn the commands for them all. The most common ones you are likely to use are shown on this page: http://www.sunilpatel.co.uk/latex-type/latex-math-symbols/

You can use this page as a reference or crib sheet, the symbols are rendered as large, high quality images so you can quickly find the LaTeX command for the symbol you need.

### 1.2.4 LaTeX on a Mac

The LaTeX distribution is available for many systems including Windows, Linux and Mac OS X. The package for OS X is called MacTeX and it contains all the applications you need – bundled together and pre-customized – for a fully working LaTeX environment and work flow.

MacTeX includes a custom dedicated LaTeX editor called TeXShop for writing your '`.tex`' files and BibDesk: a program to manage your references and create your bibliography section just as easily as managing songs and creating playlists in iTunes.

## 1.3 Getting Started with this Template

If you are familiar with LaTeX, then you should explore the directory structure of the template and then proceed to place your own information into the *THESIS INFORMATION* block of the `main.tex` file. You can then modify the rest of this file to your unique specifications based on your degree/university. Section **??** on page **??** will help you do this. Make sure you also read section **??** about thesis conventions to get the most out of this template.

If you are new to LaTeX it is recommended that you carry on reading through the rest of the information in this document.

Before you begin using this template you should ensure that its style complies with the thesis style guidelines imposed by your institution. In most cases this template style and layout will be suitable. If it is not, it may only require a small change to bring the template in line with your institution's recommendations. These modifications will need to be done on the `MastersDoctoralThesis.cls` file.

### 1.3.1 About this Template

This LaTeX Thesis Template is originally based and created around a LaTeX style file created by Steve R. Gunn from the University of Southampton (UK), department of Electronics and Computer Science. You can find his original thesis style file at his site, here: http://www.ecs.soton.ac.uk/~srg/softwaretools/document/templates/

Steve's **ecsthesis.cls** was then taken by Sunil Patel who modified it by creating a skeleton framework and folder structure to place the thesis files in. The resulting template can be found on Sunil's site here: http://www.sunilpatel.co.uk/thesis-template

Sunil's template was made available through http://www.LaTeXTemplates.com where it was modified many times based on user requests and questions. Version 2.0 and onwards of this template represents a major modification to Sunil's template and is, in fact, hardly recognisable. The work to make version 2.0 possible was carried out by Vel and Johannes Böttcher.

## 1.4 What this Template Includes

### 1.4.1 Folders

This template comes as a single zip file that expands out to several files and folders. The folder names are mostly self-explanatory:

**Appendices** – this is the folder where you put the appendices. Each appendix should go into its own separate **.tex** file. An example and template are included in the directory.

**Chapters** – this is the folder where you put the thesis chapters. A thesis usually has about six chapters, though there is no hard rule on this. Each chapter should go in its own separate **.tex** file and they can be split as:

- Chapter 1: Introduction to the thesis topic
- Chapter 2: Background information and theory
- Chapter 3: (Laboratory) experimental setup
- Chapter 4: Details of experiment 1
- Chapter 5: Details of experiment 2
- Chapter 6: Discussion of the experimental results
- Chapter 7: Conclusion and future directions

This chapter layout is specialised for the experimental sciences, your discipline may be different.

**Figures** – this folder contains all figures for the thesis. These are the final images that will go into the thesis document.

### 1.4.2 Files

Included are also several files, most of them are plain text and you can see their contents in a text editor. After initial compilation, you will see that more auxiliary files are created by LaTeX or BibTeX and which you don't need to delete or worry about:

**example.bib** – this is an important file that contains all the bibliographic information and references that you will be citing in the thesis for use with BibTeX. You can write it manually, but there are reference manager programs available that will create and manage it for you. Bibliographies in LaTeX are a large subject and you may need to read about BibTeX before starting with this. Many modern reference

managers will allow you to export your references in BibTeX format which greatly eases the amount of work you have to do.

**MastersDoctoralThesis.cls** – this is an important file. It is the class file that tells LaTeX how to format the thesis.

**main.pdf** – this is your beautifully typeset thesis (in the PDF file format) created by LaTeX. It is supplied in the PDF with the template and after you compile the template you should get an identical version.

**main.tex** – this is an important file. This is the file that you tell LaTeX to compile to produce your thesis as a PDF file. It contains the framework and constructs that tell LaTeX how to layout the thesis. It is heavily commented so you can read exactly what each line of code does and why it is there. After you put your own information into the *THESIS INFORMATION* block – you have now started your thesis!

Files that are *not* included, but are created by LaTeX as auxiliary files include:

**main.aux** – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file.

**main.bbl** – this is an auxiliary file generated by BibTeX, if it is deleted, BibTeX simply regenerates it when you run the `main.aux` file. Whereas the `.bib` file contains all the references you have, this `.bbl` file contains the references you have actually cited in the thesis and is used to build the bibliography section of the thesis.

**main.blg** – this is an auxiliary file generated by BibTeX, if it is deleted BibTeX simply regenerates it when you run the main `.aux` file.

**main.lof** – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file. It tells LaTeX how to build the *List of Figures* section.

**main.log** – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file. It contains messages from LaTeX, if you receive errors and warnings from LaTeX, they will be in this `.log` file.

**main.lot** – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file. It tells LaTeX how to build the *List of Tables* section.

**main.out** – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file.

So from this long list, only the files with the `.bib`, `.cls` and `.tex` extensions are the most important ones. The other auxiliary files can be ignored or deleted as LaTeX and BibTeX will regenerate them.

## 1.5 Filling in Your Information in the `main.tex` File

You will need to personalise the thesis template and make it your own by filling in your own information. This is done by editing the `main.tex` file in a text editor or your favourite LaTeX environment.

Open the file and scroll down to the third large block titled *THESIS INFORMATION* where you can see the entries for *University Name*, *Department Name*, etc ...

Fill out the information about yourself, your group and institution. You can also insert web links, if you do, make sure you use the full URL, including the `http://` for this. If you don't want these to be linked, simply remove the `\href{url}{name}` and only leave the name.

When you have done this, save the file and recompile `main.tex`. All the information you filled in should now be in the PDF, complete with web links. You can now begin your thesis proper!

## 1.6 The `main.tex` File Explained

The **`main.tex`** file contains the structure of the thesis. There are plenty of written comments that explain what pages, sections and formatting the LaTeX code is creating. Each major document element is divided into commented blocks with titles in all capitals to make it obvious what the following bit of code is doing. Initially there seems to be a lot of LaTeX code, but this is all formatting, and it has all been taken care of so you don't have to do it.

Begin by checking that your information on the title page is correct. For the thesis declaration, your institution may insist on something different than the text given. If this is the case, just replace what you see with what is required in the *DECLARATION PAGE* block.

Then comes a page which contains a funny quote. You can put your own, or quote your favourite scientist, author, person, and so on. Make sure to put the name of the person who you took the quote from.

Following this is the abstract page which summarises your work in a condensed way and can almost be used as a standalone document to describe what you have done. The text you write will cause the heading to move up so don't worry about running out of space.

Next come the acknowledgements. On this page, write about all the people who you wish to thank (not forgetting parents, partners and your advisor/supervisor).

The contents pages, list of figures and tables are all taken care of for you and do not need to be manually created or edited. The next set of pages are more likely to be optional and can be deleted since they are for a more technical thesis: insert a list of abbreviations you have used in the thesis, then a list of the physical constants and numbers you refer to and finally, a list of mathematical symbols used in any formulae. Making the effort to fill these tables means the reader has a one-stop place to refer to instead of searching the internet and references to try and find out what you meant by certain abbreviations or symbols.

The list of symbols is split into the Roman and Greek alphabets. Whereas the abbreviations and symbols ought to be listed in alphabetical order (and this is *not* done automatically for you) the list of physical constants should be grouped into similar themes.

The next page contains a one line dedication. Who will you dedicate your thesis to?

Finally, there is the block where the chapters are included. Uncomment the lines (delete the `%` character) as you write the chapters. Each chapter should be written in its own file and put into the *Chapters* folder and named **`Chapter1`**, **`Chapter2`**, etc... Similarly for the appendices, uncomment the lines as you need them. Each appendix should go into its own file and placed in the *Appendices* folder.

After the preamble, chapters and appendices finally comes the bibliography. The bibliography style (called ***`authoryear`***) is used for the bibliography and is a fully featured style that will even include links to where the referenced paper can be found online. Do not underestimate how grateful your reader will be to find that a reference to a paper is just a click away. Of course, this relies on you putting the URL information into the BibTeX file in the first place.

## 1.7 Thesis Features and Conventions

To get the best out of this template, there are a few conventions that you may want to follow.

One of the most important (and most difficult) things to keep track of in such a long document as a thesis is consistency. Using certain conventions and ways of doing things (such as using a Todo list) makes the job easier. Of course, all of these are optional and you can adopt your own method.

### 1.7.1 Printing Format

This thesis template is designed for double sided printing (i.e. content on the front and back of pages) as most theses are printed and bound this way. Switching to one sided printing is as simple as uncommenting the *oneside* option of the `documentclass` command at the top of the **`main.tex`** file. You may then wish to adjust the margins to suit specifications from your institution.

The headers for the pages contain the page number on the outer side (so it is easy to flick through to the page you want) and the chapter name on the inner side.

The text is set to 11 point by default with single line spacing, again, you can tune the text size and spacing should you want or need to using the options at the very start of **`main.tex`**. The spacing can be changed similarly by replacing the *singlespacing* with *onehalfspacing* or *doublespacing*.

### 1.7.2 Using US Letter Paper

The paper size used in the template is A4, which is the standard size in Europe. If you are using this thesis template elsewhere and particularly in the United States, then you may have to change the A4 paper size to the US Letter size. This can be done in the margins settings section in **`main.tex`**.

Due to the differences in the paper size, the resulting margins may be different to what you like or require (as it is common for institutions to dictate certain margin sizes). If this is the case, then the margin sizes can be tweaked by modifying the values in the same block as where you set the paper size. Now your document should be set up for US Letter paper size with suitable margins.

### 1.7.3 References

The `biblatex` package is used to format the bibliography and inserts references such as this one (**Reference1**). The options used in the **`main.tex`** file mean that the in-text citations of references are formatted with the author(s) listed with the date of the publication. Multiple references are separated by semicolons (e.g. (**Reference2**; **Reference1**)) and references with more than three authors only show the first author with *et al.* indicating there are more authors (e.g. (**Reference3**)). This is done automatically for you. To see how you use references, have a look at the **`Chapter1.tex`** source file. Many reference managers allow you to simply drag the reference into the document as you type.

Scientific references should come *before* the punctuation mark if there is one (such as a comma or period). The same goes for footnotes[1]. You can change this but the most important thing is to keep the convention consistent throughout the thesis. Footnotes themselves should be full, descriptive sentences (beginning with a capital

[1] Such as this footnote, here down at the bottom of the page.

letter and ending with a full stop). The APA6 states: "Footnote numbers should be superscripted, [...], following any punctuation mark except a dash." The Chicago manual of style states: "A note number should be placed at the end of a sentence or clause. The number follows any punctuation mark except the dash, which it precedes. It follows a closing parenthesis."

The bibliography is typeset with references listed in alphabetical order by the first author's last name. This is similar to the APA referencing style. To see how LaTeX typesets the bibliography, have a look at the very end of this document (or just click on the reference number links in in-text citations).

#### A Note on bibtex

The bibtex backend used in the template by default does not correctly handle unicode character encoding (i.e. "international" characters). You may see a warning about this in the compilation log and, if your references contain unicode characters, they may not show up correctly or at all. The solution to this is to use the biber backend instead of the outdated bibtex backend. This is done by finding this in **main.tex**: *backend=bibtex* and changing it to *backend=biber*. You will then need to delete all auxiliary BibTeX files and navigate to the template directory in your terminal (command prompt). Once there, simply type `biber main` and biber will compile your bibliography. You can then compile **main.tex** as normal and your bibliography will be updated. An alternative is to set up your LaTeX editor to compile with biber instead of bibtex, see here for how to do this for various editors.

### 1.7.4 Tables

Tables are an important way of displaying your results, below is an example table which was generated with this code:

```
\begin{table}
\caption{The effects of treatments X and Y on the four groups studied.}
\label{tab:treatments}
\centering
\begin{tabular}{l l l}
\toprule
\tabhead{Groups} & \tabhead{Treatment X} & \tabhead{Treatment Y} \\
\midrule
1 & 0.2 & 0.8\\
2 & 0.17 & 0.7\\
3 & 0.24 & 0.75\\
4 & 0.68 & 0.3\\
\bottomrule\\
\end{tabular}
\end{table}
```

You can reference tables with `\ref{<label>}` where the label is defined within the table environment. See **Chapter1.tex** for an example of the label and citation (e.g. Table **??**).

### 1.7.5 Figures

There will hopefully be many figures in your thesis (that should be placed in the *Figures* folder). The way to insert figures into your thesis is to use a code template like this:

TABLE 1.1: The effects of treatments X and Y on the four groups studied.

| Groups | Treatment X | Treatment Y |
|---|---|---|
| 1 | 0.2 | 0.8 |
| 2 | 0.17 | 0.7 |
| 3 | 0.24 | 0.75 |
| 4 | 0.68 | 0.3 |

```
\begin{figure}
\centering
\includegraphics{Figures/Electron}
\decoRule
\caption[An Electron]{An electron (artist's impression).}
\label{fig:Electron}
\end{figure}
```

Also look in the source file. Putting this code into the source file produces the picture of the electron that you can see in the figure below.

FIGURE 1.1: An electron (artist's impression).

Sometimes figures don't always appear where you write them in the source. The placement depends on how much space there is on the page for the figure. Sometimes there is not enough room to fit a figure directly where it should go (in relation to the text) and so LaTeX puts it at the top of the next page. Positioning figures is the job of LaTeX and so you should only worry about making them look good!

Figures usually should have captions just in case you need to refer to them (such as in Figure **??**). The `\caption` command contains two parts, the first part, inside the square brackets is the title that will appear in the *List of Figures*, and so should be short. The second part in the curly brackets should contain the longer and more descriptive caption text.

The `\decoRule` command is optional and simply puts an aesthetic horizontal line below the image. If you do this for one image, do it for all of them.

LaTeX is capable of using images in pdf, jpg and png format.

### 1.7.6 Typesetting mathematics

If your thesis is going to contain heavy mathematical content, be sure that LaTeX will make it look beautiful, even though it won't be able to solve the equations for you.

The "Not So Short Introduction to LaTeX" (available on CTAN) should tell you everything you need to know for most cases of typesetting mathematics. If you need more information, a much more thorough mathematical guide is available from the AMS called, "A Short Math Guide to LaTeX" and can be downloaded from: `ftp://ftp.ams.org/pub/tex/doc/amsmath/short-math-guide.pdf`

There are many different LaTeX symbols to remember, luckily you can find the most common symbols in The Comprehensive LaTeX Symbol List.

You can write an equation, which is automatically given an equation number by LaTeX like this:

```
\begin{equation}
E = mc^{2}
\label{eqn:Einstein}
\end{equation}
```

This will produce Einstein's famous energy-matter equivalence equation:

$$E = mc^2 \tag{1.1}$$

All equations you write (which are not in the middle of paragraph text) are automatically given equation numbers by LaTeX. If you don't want a particular equation numbered, use the unnumbered form:

```
\[ a^{2}=4 \]
```

## 1.8 Sectioning and Subsectioning

You should break your thesis up into nice, bite-sized sections and subsections. LaTeX automatically builds a table of Contents by looking at all the `\chapter{}`, `\section{}` and `\subsection{}` commands you write in the source.

The Table of Contents should only list the sections to three (3) levels. A `chapter{}` is level zero (0). A `\section{}` is level one (1) and so a `\subsection{}` is level two (2). In your thesis it is likely that you will even use a `subsubsection{}`, which is level three (3). The depth to which the Table of Contents is formatted is set within `MastersDoctoralThesis.cls`. If you need this changed, you can do it in `main.tex`.

## 1.9 In Closing

You have reached the end of this mini-guide. You can now rename or overwrite this pdf file and begin writing your own `Chapter1.tex` and the rest of your thesis. The

easy work of setting up the structure and framework has been taken care of for you. It's now your job to fill it out!

Good luck and have lots of fun!

Guide written by —
Sunil Patel: www.sunilpatel.co.uk
Vel: LaTeXTemplates.com

# Chapter 2

# A Probabilistic Graph Coupling View of Dimension Reduction

## 2.1 Introduction

Dimensionality reduction (DR) is of central importance when dealing with high-dimensional data (**donoho2000high**). It mitigates the curse of dimensionality, allowing for greater statistical flexibility and less computational complexity. DR also enables visualization that can be of great practical interest for understanding and interpreting the structure of large datasets. Most seminal approaches include Principal Component Analysis (PCA) **pearson1901liii**, multidimensional scaling **kruskal1978multidimensional** and more broadly kernel eigenmaps methods such as Isomap **balasubramanian2002isomap**, Laplacian eigenmaps (**belkin2003laplacian**) and diffusion maps (**coifman2006diffusion**). These methods share the definition of a pairwise similarity kernel that assigns a high value to close neighbors and the resolution of a spectral problem. They are well understood and unified in the kernel PCA framework (**ham2004kernel**).

In the past decade, the field has witnessed a major shift with the emergence of a new class of methods. They are also based on pairwise similarities but these are not converted into inner products. Instead, they define pairwise similarity functions in both input and latent spaces and optimize a cost between the two. Among such methods, the Stochastic Neighbor Embedding (SNE) algorithm **NIPS2002SNE**, its heavy-tailed symmetrized version t-SNE **maaten2008tSNE** or more recent approaches like LargeVis **tang2016visualizing** and UMAP **mcinnes2018umap** are arguably the most used in practice. These will be referred to as *SNE-like* or *neighbor embedding* methods in what follows. They are increasingly popular and now considered as the state-of-art techniques in many fields **li2017application**; **kobak2019art**; **anders2018dissecting**. Their popularity is mainly due to their exceptional ability to preserve local structure, *i.e.* close points in the input space have close embeddings, as shown empirically **wang2021understanding**. They also demonstrate impressive performances in identifying clusters **arora2018analysis**; **linderman2019clustering**. However this is done at the expense of global structure, that these methods struggle in preserving **wattenberg2016use**; **coenen2019understanding** *i.e.* the relative large-scale distances between embedded points do not necessarily correspond to the original ones.

Due to a lack of clear probabilistic foundations, these properties remain mostly empirical. This gap between theory and practice is detrimental as practitioners may rely on strategies that are not optimal for their use case. While recent software developments are making these methods more scalable **chan2018t**; **pezzotti2019gpgpu**; **linderman2019fast** and further expanding their use, the need for a well-established probabilistic framework is becoming more prominent. In this work we define the

generative probabilistic model that encompasses current embedding methods, while establishing new links with the well-established PCA model.

**Outline.** Consider $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_n)^\top \in \mathbb{R}^{n \times p}$, an input dataset that consists of $n$ vectors of dimension $p$. Our task is to embed $\mathbf{X}$ in a lower dimensional space of dimension $q < p$ (typically $q = 2$ for visualization), and we denote by $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_n)^\top \in \mathbb{R}^{n \times q}$ the unknown embeddings. The rationale of our framework is to suppose that the observations $\mathbf{X}$ and $\mathbf{Z}$ are structured by two latent graphs with $\mathbf{W}_X$ and $\mathbf{W}_Z$ standing for their $n$-square weight matrices. As the goal of DR is to preserve the input's structure in the latent space, we propose to find the best low-dimensional representation $\mathbf{Z}$ of $\mathbf{X}$ such that $\mathbf{W}_X$ and $\mathbf{W}_Z$ are close. To build a flexible and robust probabilistic framework, we consider random graphs distributed according to some predefined prior distributions. Our objective is to match the posterior distributions of $\mathbf{W}_X$ and $\mathbf{W}_Z$. Note that as they share the same dimensionality the latter graphs can be easily compared unlike $\mathbf{X}$ and $\mathbf{Z}$. The coupling is done with a cross entropy criterion, the minimization of which will be referred to as graph coupling.

In this work, our main contributions are as follows.

- We show that SNE, t-SNE, LargeVis and UMAP are all instances of graph coupling and characterized by different choices of prior for discrete latent structuring graphs (**??**). We demonstrate that such graphs essentially capture conditional independencies among rows through a pairwise Markov Random Field (MRF) model which construction can be found in **??**.
- We uncover the intrinsic probabilistic property explaining why such methods perform poorly on conserving the large scale structure of the data as a consequence of a degeneracy of the MRF when shift invariant kernels are used (**??**). Such degeneracy induces the loss of the relative positions of clusters corresponding to the connected components of the posterior latent graphs which distributions are identified (**??**). These findings are highlighted by a new initialization of the embeddings that is empirically tested (**??**).
- We show that for Gaussian MRFs, when adapting graph coupling to precision matrices with suitable priors, PCA appears as a natural extension of the coupling problem in its continuous version (**??**). Such model does not suffer from the aforementioned degeneracy hence preserves the large-scale structure.

## 2.2 Shift-Invariant Pairwise MRF to Model Row Dependencies

We start by defining the distribution of the observations given a graph. The latter takes the form of a pairwise MRF model which as we show is improper (*i.e.* not integrable on $\mathbb{R}^{n \times p}$) when shift-invariant kernels are used. We consider a fixed directed graph $\mathbf{W} \in \mathcal{S}_W$ where:

$$\mathcal{S}_W = \left\{ \mathbf{W} \in \mathbb{N}^{n \times n} \mid \forall (i,j) \in [n]^2, W_{ii} = 0, W_{ij} \leq n \right\}$$

Throughout, $(E, \mathcal{B}(E), \lambda_E)$ denotes a measure space where $\mathcal{B}(E)$ is the Borel $\sigma$-algebra on $E$ and $\lambda_E$ is the Lebesgue measure on $E$.

### 2.2.1 Graph Laplacian Null Space

A central element in our construction is the graph Laplacian linear map, defined as follows, where $\mathcal{S}_+^n(\mathbb{R})$ is the set of positive semidefinite matrices.

**Definition 1.** The graph Laplacian operator is the map $L\colon \mathbb{R}_+^{n\times n} \to \mathcal{S}_+^n(\mathbb{R})$ such that

$$\text{for } (i,j)\in[n]^2, \quad L(\mathbf{W})_{ij} = \begin{cases} -W_{ij} & \text{if } i \neq j \\ \sum_{k\in[n]} W_{ik} & \text{otherwise}\,. \end{cases}$$

With an abuse of notation, let $\boldsymbol{L} = L(\overline{\mathbf{W}})$ where $\overline{\mathbf{W}} = \mathbf{W} + \mathbf{W}^\top$. Let $(C_1, ..., C_R)$ be a partition of $[n]$ (*i.e.* the set $\{1, 2, ..., n\}$) corresponding to the connected components (CCs) of $\overline{\mathbf{W}}$. As well known in spectral graph theory **Chung97**, the null space of $\boldsymbol{L}$ is spanned by the orthonormal vectors $\{\boldsymbol{U}_r\}_{r\in[R]}$ such that for $r \in [R]$, $\boldsymbol{U}_r = \left(n_r^{-1/2}\mathbb{1}_{i\in C_r}\right)_{i\in[n]}$ with $n_r = \mathrm{Card}(C_r)$. By the spectral theorem, $\boldsymbol{U}_{[R]}$ can be completed such that $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ where $\boldsymbol{U} = (\boldsymbol{U}_1, ..., \boldsymbol{U}_n)$ is orthogonal and $\boldsymbol{\Lambda} = \mathrm{diag}((\lambda_i)_{i\in[n]})$ with $0 = \lambda_1 = ... = \lambda_R < \lambda_{R+1} \leq ... \leq \lambda_n$. In the following, the data is split into two parts: $\mathbf{X}_M$, the orthogonal projection of $\mathbf{X}$ on $\mathcal{S}_M = (\ker \boldsymbol{L}) \otimes \mathbb{R}^p$, and $\mathbf{X}_C$, the projection on $\mathcal{S}_C = (\ker \boldsymbol{L})^\perp \otimes \mathbb{R}^p$. For $i \in [n]$, $\mathbf{X}_{M,i} = \sum_{r\in[R]} n_r^{-1}\mathbb{1}_{i\in C_r}\sum_{\ell\in C_r}\mathbf{X}_\ell$ hence $\mathbf{X}_M$ stands for the empirical means of $\mathbf{X}$ on CCs, thus modelling the CC positions, while $\mathbf{X}_C = \mathbf{X} - \mathbf{X}_M$ is CC-wise centered, thus modeling the relative positions of the nodes within CCs. We now introduce the probability distribution of these variables.

### 2.2.2 Pairwise MRF and Shift-Invariances

In this work, the dependency structure among rows of the data is governed by a graph. The strength of the connection between two nodes is given by a symmetric function $k : \mathbb{R}^p \to \mathbb{R}_+$. We consider the following pairwise MRF unnormalized density function:

$$f_k\colon (\mathbf{X}, \mathbf{W}) \mapsto \prod_{(i,j)\in[n]^2} k(\mathbf{X}_i - \mathbf{X}_j)^{W_{ij}}\,. \tag{2.1}$$

As we will see shortly, the above is at the heart of DR methods based on pairwise similarities. Note that as $k$ measures the similarity between couples of samples, $f_k$ will take high values if the rows of $\mathbf{X}$ vary smoothly on the graph $\mathbf{W}$. Thus we can expect $\mathbf{X}_i$ and $\mathbf{X}_j$ to be close if there is an edge between node $i$ and node $j$ in $\mathbf{W}$. A key remark is that $f_k$ is kept invariant by translating $\mathbf{X}_M$. Namely for all $\mathbf{X} \in \mathbb{R}^{n\times p}$, $f_k(\mathbf{X}, \mathbf{W}) = f_k(\mathbf{X}_C, \mathbf{W})$. This invariance results in $f_k(\cdot, \mathbf{W})$ being non integrable on $\mathbb{R}^{n\times p}$, as we see with the following example.

**Gaussian kernel.** For a positive definite matrix $\boldsymbol{\Sigma} \in \mathcal{S}_{++}^n(\mathbb{R})$, consider the Gaussian kernel $k : \boldsymbol{x} \mapsto e^{-\frac{1}{2}\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}}^2}$ where $\boldsymbol{\Sigma}$ stands for the covariance among columns. One has:

$$\log f_k(\mathbf{X}, \mathbf{W}) = -\sum_{(i,j)\in[n]^2} W_{ij}\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_{\boldsymbol{\Sigma}}^2 = -\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^T\boldsymbol{L}\boldsymbol{X}\right) \tag{2.2}$$

by property of the graph Laplacian (**??**). In this case, it is clear that due to the rank deficiency of $\boldsymbol{L}$, $f_k(\cdot, \mathbf{W})$ is only $\lambda_{\mathcal{S}_C}$-integrable. In general DR settings one does not want to rely on Gaussian kernels only. A striking example is the use of the Student kernel in t-SNE **maaten2008tSNE**. Heavy-tailed kernels appear useful

when the dimension of the embeddings is smaller than the intrinsic dimension of the data **kobak2019heavy**. Our contribution provides flexibility by extending the previous result to a large class of kernels, as stated in the following theorem.

**Theorem 2.** *If $k$ is $\lambda_{\mathbb{R}^p}$-integrable and bounded above $\lambda_{\mathbb{R}^p}$-almost everywhere then $f_k(\cdot, \mathbf{W})$ is $\lambda_{\mathcal{S}_C}$-integrable.*

We refer to **??** for the proof. We can now define a distribution on $(\mathcal{S}_C, \mathcal{B}(\mathcal{S}_C))$, where $\mathcal{C}_k(\mathbf{W}) = \int f_k(\cdot, \mathbf{W}) d\lambda_{\mathcal{S}_C}$:

$$\mathbb{P}_k(d\mathbf{X}_C|\mathbf{W}) = \mathcal{C}_k(\mathbf{W})^{-1} f_k(\mathbf{X}_C, \mathbf{W}) \lambda_{\mathcal{S}_C}(d\mathbf{X}_C)\,. \tag{2.3}$$

**Remark 3.** Kernels may have node-specific bandwidths $\boldsymbol{\tau}$, set during a pre-processing step, giving $f_k(\mathbf{X}, \mathbf{W}) = \prod_{(i,j)} k((\mathbf{X}_i - \mathbf{X}_j)/\tau_i)^{W_{ij}}$. Note that such bandwidth does not affect the degeneracy of the distribution and **??** still holds.

**Between-Rows Dependency Structure.** By symmetry of $k$, reindexing gives: $f_k(\mathbf{X}, \mathbf{W}) = \prod_{j\in[n]} \prod_{i\in[j]} k(\mathbf{X}_i - \mathbf{X}_j)^{\overline{W}_{ij}}$. Hence distribution (**??**) boils down to a pairwise MRF model (**clifford1990markov**) with respect to the undirected graph $\overline{\mathbf{W}}$, $\mathcal{C}_k$ playing the role of the partition function. Note that since $f_k$ (Equation **??**) trivially factorize according to the cliques of $\overline{\mathbf{W}}$, the Hammersley-Clifford theorem ensures that the rows of $\mathbf{X}_C$ satisfy the local and global Markov properties with respect to $\overline{\mathbf{W}}$.

### 2.2.3 Uninformative Model for CC-wise Means

We showed that the MRF (**??**) is only integrable on $\mathcal{S}_C$, the definition of which depends on the connectivity structure of $\mathbf{W}$. As we now demonstrate, the latter MRF can be seen as a limit of proper distributions on $\mathbb{R}^{n\times p}$, see *e.g.* **rue2005gaussian** for a similar construction in the Gaussian case. We introduce the Borel function $f^\varepsilon(\cdot, \mathbf{W})\colon \mathbb{R}^{n\times p} \to \mathbb{R}_+$ for $\varepsilon > 0$ such that for all $\mathbf{X} \in \mathbb{R}^{n\times p}$, $f^\varepsilon(\mathbf{X}, \mathbf{W}) = f^\varepsilon(\mathbf{X}_M, \mathbf{W})$. To allow $f^\varepsilon$ to become arbitrarily non-informative, we assume that for all $\mathbf{W} \in \mathcal{S}_W$, $f^\varepsilon(\cdot, \mathbf{W})$ is $\lambda_{\mathcal{S}_M}$-integrable for all $\varepsilon \in \mathbb{R}_+^*$ and $f^\varepsilon(\cdot, \mathbf{W}) \xrightarrow[\varepsilon\to 0]{} 1$ almost everywhere. We now define the conditional distribution on $(\mathcal{S}_M, \mathcal{B}(\mathcal{S}_M))$ as follows:

$$\mathbb{P}^\varepsilon(d\mathbf{X}_M|\mathbf{W}) = \mathcal{C}^\varepsilon(\mathbf{W})^{-1} f^\varepsilon(\mathbf{X}_M, \mathbf{W}) \lambda_{\mathcal{S}_M}(d\mathbf{X}_M) \tag{2.4}$$

where $\mathcal{C}^\varepsilon(\mathbf{W}) = \int f^\varepsilon(\cdot, \mathbf{W}) d\lambda_{\mathcal{S}_M}$. With this at hand, the joint conditional is defined as the product measure of (**??**) and (**??**) over the row axis, the integrability of which is ensured by the Fubini-Tonelli theorem. In the following we will use the compact notation $\mathcal{C}_k^\varepsilon(\mathbf{W}) = \mathcal{C}_k(\mathbf{W})\mathcal{C}^\varepsilon(\mathbf{W})$ for the joint normalizing constant.

**Remark 4.** At the limit $\varepsilon \to 0$ the above construction amounts to setting an infinite variance on the distribution of the empirical means of $\mathbf{X}$ on CCs, thus loosing the inter-CC structure.

As an illustration, one can structure the CCs' relative positions according to a Gaussian model with positive definite precision $\varepsilon\boldsymbol{\Theta} \in \mathcal{S}_{++}^R(\mathbb{R})$, as it amounts to choosing $f^\varepsilon : \mathbf{X} \to \exp\left(-\frac{\varepsilon}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{X}^\top \boldsymbol{U}_{[:R]}\boldsymbol{\Theta}\boldsymbol{U}_{[R]}^\top \mathbf{X}\right)\right)$ such that: $\operatorname{vec}(\mathbf{X}_M)|\boldsymbol{\Theta} \sim \mathcal{N}\left(\mathbf{0}, \left(\varepsilon \boldsymbol{U}_{[:R]}\boldsymbol{\Theta}\boldsymbol{U}_{[R]}^\top\right)^{-1} \otimes \boldsymbol{\Sigma}\right)$ where $\otimes$ denotes the Kronecker product.

## 2.3 Graph Coupling as a Unified Objective for Pairwise Similarity Methods

In this section, we show that neighbor embedding methods can be recovered in the presented framework. They are obtained, for particular choices of graph priors, at the limit $\varepsilon \to 0$ when $f^\varepsilon$ becomes non informative and the CCs' relative positions are lost.

We now turn to the priors for $\mathbf{W}$. Our methodology is similar to that of constructing conjugate priors for distributions in the exponential family **wainwright2008graphical**, notably we insert the cumulant function $\mathcal{C}_k^\varepsilon$ (*i.e.* normalizing constant of the conditional) as a multivariate term of the prior. We consider different forms: binary ($B$), unitary out-degree ($D$) and $n$-edges ($E$), relying on an additional term ($\Omega$) to constraint the topology of the graph. For a matrix $\boldsymbol{A}$, $A_{i+}$ denotes $\sum_j A_{ij}$ and $A_{++}$ denotes $\sum_{ij} A_{ij}$. In the following, $\boldsymbol{\pi}$ plays the role of the edge's prior. The latter can be leveraged to incorporate some additional information about the dependency structure, for instance when a network is observed **li2020high**.

**Definition 5.** Let $\boldsymbol{\pi} \in \mathbb{R}_+^{n\times n}$, $\varepsilon \in \mathbb{R}_+$, $\alpha \in \mathbb{R}$, $k$ satisfies the assumptions of **??** and $\mathcal{P} \in \{B, D, E\}$. For $\mathbf{W} \in \mathcal{S}_W$ we introduce:

$$\mathbb{P}^\varepsilon_{\mathcal{P},k}(\boldsymbol{W}; \boldsymbol{\pi}, \alpha) \propto \mathcal{C}_k^\varepsilon(\mathbf{W})^\alpha \, \Omega_\mathcal{P}(\mathbf{W}) \prod_{(i,j)\in[n]^2} \pi_{ij}^{W_{ij}}$$

where $\Omega_B(\mathbf{W}) = \prod_{ij} \mathbb{1}_{W_{ij}\leq 1}$, $\Omega_D(\mathbf{W}) = \prod_i \mathbb{1}_{W_{i+}=1}$ and $\Omega_E(\mathbf{W}) = \mathbb{1}_{W_{++}=n} \prod_{ij}(W_{ij}!)^{-1}$.

When $\alpha = 0$, the above no longer depends on $\varepsilon$ and $k$. We will use the compact notation $\mathbb{P}_\mathcal{P}(\mathbf{W}; \boldsymbol{\pi}) = \mathbb{P}^\varepsilon_{\mathcal{P},k}(\boldsymbol{W}; \boldsymbol{\pi}, 0)$. Note that by $\mathbf{W} \sim \mathbb{P}_\mathcal{P}(\cdot\,; \boldsymbol{\pi})$ we have the following simple Bernoulli ($\mathcal{B}$) and multinomial ($\mathcal{M}$) distributions, where matrix or vector division is to be understood as element-wise.

- If $\mathcal{P} = B$, $\forall (i,j) \in [n]^2$, $W_{ij} \overset{\perp\!\!\!\perp}{\sim} \mathcal{B}\left(\pi_{ij}/(1+\pi_{ij})\right)$.
- If $\mathcal{P} = D$, $\forall i \in [n]$, $\mathbf{W}_i \overset{\perp\!\!\!\perp}{\sim} \mathcal{M}\left(1, \boldsymbol{\pi}_i/\pi_{i+}\right)$.
- If $\mathcal{P} = E$, $\mathbf{W} \sim \mathcal{M}\left(n, \boldsymbol{\pi}/\pi_{++}\right)$.

We now show that the posterior distribution of the graph given the observations takes a simple form when the distribution of CC empirical means $\boldsymbol{X}_M$ diffuses *i.e.* when $\varepsilon \to 0$ (a proof of the following result can be found in **??**). In the following, $\odot$ stands for the Hadamard product and $\mathcal{D}$ for the convergence in distribution.

**Proposition 6.** *Let $\boldsymbol{\pi} \in \mathbb{R}_+^{n\times n}$, $k$ satisfies the assumptions of* **??** *with $\boldsymbol{K}_X = (k(\mathbf{X}_i - \mathbf{X}_j))_{(i,j)\in[n]^2}$ and $\mathcal{P} \in \{B, D, E\}$. If $\mathbf{W}^\varepsilon \sim \mathbb{P}^\varepsilon_{\mathcal{P},k}(\cdot\,; \boldsymbol{\pi}, 1)$ then*

$$\mathbf{W}^\varepsilon | \mathbf{X} \xrightarrow[\varepsilon\to 0]{\mathcal{D}} \mathbb{P}_\mathcal{P}(\cdot\,; \boldsymbol{\pi} \odot \boldsymbol{K}_X)\,.$$

**Remark 7.** For all $\mathbf{W} \in \mathcal{S}_W$, $\mathcal{C}^\varepsilon(\mathbf{W})$ diverges as $\varepsilon \to 0$, hence the graph prior (**??**) is improper at the limit. This compensates for the uninformative diffuse conditional and allows to retrieve a well-defined tractable posterior limit.

### 2.3.1 Retrieving Well Known DR Methods

We now provide a unified view of neighbor embedding objectives as a coupling between graph posterior distributions. To that extent we derive the cross entropy associated with the various graph priors at hand. In what follows, $k_x$ and $k_z$ satisfy

the assumptions of **??** and we denote by $\boldsymbol{K}_X$ and $\boldsymbol{K}_Z$ the associated kernel matrices on $\mathbf{X}$ and $\mathbf{Z}$ respectively. For both graph priors we consider the parameters $\boldsymbol{\pi} = \mathbf{1}$ and $\alpha = 1$. For $(\mathcal{P}_X, \mathcal{P}_Z) \in \{B, D, E\}^2$, we introduce the cross entropy between the limit posteriors at $\varepsilon \to 0$,

$$\mathcal{H}_{\mathcal{P}_X,\mathcal{P}_Z} = -\mathbb{E}_{\mathbf{W}_X \sim \mathbb{P}_{\mathcal{P}_X}(\cdot;\boldsymbol{K}_X)}[\log \mathbb{P}_{\mathcal{P}_Z}(\mathbf{W}_Z = \mathbf{W}_X; \boldsymbol{K}_Z)]$$

defining a coupling criterion to be optimized with respect to embedding coordinates $\mathbf{Z}$. We now go through each couple $(\mathcal{P}_X, \mathcal{P}_Z)$ such that $\text{supp}\,(\mathbb{P}_{\mathcal{P}_X}) \subset \text{supp}\,(\mathbb{P}_{\mathcal{P}_Z})$ for the cross-entropy to be defined.

**SNE.** When $\mathcal{P}_X = \mathcal{P}_Z = D$, the probability of the limit posterior graphs factorizes over the nodes and the cross-entropy between limit posteriors takes the form of the objective of SNE **hinton2002stochastic**, where for $i \in [n], \boldsymbol{P}_i^D = \boldsymbol{K}_{X,i}/K_{X,i+}$ and $\boldsymbol{Q}_i^D = \boldsymbol{K}_{Z,i}/K_{Z,i+}$,

$$\mathcal{H}_{D,D} = -\sum_{i \neq j} P_{ij}^D \log Q_{ij}^D \,.$$

**Symmetric-SNE.** Choosing $\mathcal{P}_X = D$ and $\mathcal{P}_Z = E$, we define for $(i,j) \in [n]^2$, $\boldsymbol{Q}_{ij}^E = K_{Z,ij}/K_{Z,++}$ and $\overline{P}_{ij}^D = P_{ij}^D + P_{ji}^D$. The symmetry of $\boldsymbol{Q}^E$ yields:

$$\mathcal{H}_{D,E} = -\sum_{i \neq j} P_{ij}^D \log Q_{ij}^E = -\sum_{i<j} \overline{P}_{ij}^D \log Q_{ij}^E$$

and the symmetrized objective of t-SNE **maaten2008tSNE** is recovered.

**LargeVis.** Now choosing $\mathcal{P}_X = D$ and $\mathcal{P}_Z = B$, one can also notice that $\boldsymbol{Q}^B = (K_{Z,ij}/(1 + K_{Z,ij}))_{(i,j)\in[n]^2}$ is symmetric. With this at hand the limit cross-entropy reads

$$\mathcal{H}_{D,B} = -\sum_{i \neq j} P_{ij}^D \log Q_{ij}^B + \left(1 - P_{ij}^D\right) \log\left(1 - Q_{ij}^B\right) = -\sum_{i<j} \overline{P}_{ij}^D \log Q_{ij}^B + \left(2 - \overline{P}_{ij}^D\right) \log(1 - Q_{ij}^B)$$

which is the objective of LargeVis **tang2016visualizing**.

**UMAP.** Let us take $\mathcal{P}_X = \mathcal{P}_Z = B$ and consider the symmetric thresholded graph $\widetilde{\mathbf{W}}_X = \mathbb{1}_{\mathbf{W}_X + \mathbf{W}_X^\top \geq 1}$. By independence of the edges, $\widetilde{W}_{X,ij} \sim \mathcal{B}\left(\widetilde{P}_{ij}^B\right)$ where $\widetilde{P}_{ij}^B = P_{ij}^B + P_{ji}^B - P_{ij}^B P_{ji}^B$ and $\boldsymbol{P}^B = (K_{X,ij}/(1 + K_{X,ij}))_{(i,j)\in[n]^2}$. Coupling $\widetilde{\mathbf{W}}_X$ and $\mathbf{W}_Z$ gives:

$$\mathcal{H}_{\widetilde{B},B} = -2\sum_{i<j} \widetilde{P}_{ij}^B \log Q_{ij}^B + \left(1 - \widetilde{P}_{ij}^B\right) \log\left(1 - Q_{ij}^B\right)$$

which is the loss function considered in UMAP **mcinnes2018umap**, the construction of $\widetilde{\mathbf{W}}_X$ being borrowed from section 3.1 of the paper.

**Remark 8.** One can also consider $\mathcal{H}_{E,E}$ but as detailed in **maaten2008tSNE**, this criterion fails at positioning outliers and is therefore not considered. Interestingly, any other feasible combination of the presented priors relates to an existing method.

TABLE 2.1: Prior distributions for $\mathbf{W}_X$ and $\mathbf{W}_Z$ associated with the pairwise similarity coupling DR algorithms. Grey-colored boxes are such that the cross-entropy is undefined.

| | $B$ | $D$ | $E$ | |
|---|---|---|---|---|
| $\widetilde{B}$ | UMAP | | | $\mathcal{P}_X$ |
| | | | | $\mathcal{P}_Z$ |
| $D$ | LARGEVIS | SNE | T-SNE | |

### 2.3.2 Interpretations

As we have seen in **??**, SNE-like methods can all be derived from the graph coupling framework. What characterizes each of them is the choice of priors considered for the latent structuring graphs. To the best of our knowledge, the presented framework is the first that manages to unify all these DR algorithms. Such a framework opens many perspectives for improving upon current practices as we discuss in **??** and **??**. We now focus on a few insights that our work provides about the empirical performances of these methods.

**Repulsion & Attraction.** Decomposing $\mathcal{H}_{\mathcal{P}_X,\mathcal{P}_Z}$ with Bayes' rule and simplifying constant terms one has the following optimization problem:

$$\min_{\mathbf{Z}\in\mathbb{R}^{n\times q}} -\sum_{(i,j)\in[n]^2} \boldsymbol{P}_{ij}^{\mathcal{P}_X} \log k_z(\mathbf{Z}_i - \mathbf{Z}_j) + \log \mathbb{P}(\mathbf{Z}). \tag{2.5}$$

The first and second terms in **??** respectively summarize the attractive and repulsive forces of the objective. Recall from **??** that $\boldsymbol{P}^{\mathcal{P}_X}$ is the posterior expectation of $\mathbf{W}_X$. Hence in SNE-like methods, the attractive forces resume to a pairwise MRF log likelihood with respect to a graph posterior expectation given $\mathbf{X}$. For instance if $k_z$ is the Gaussian kernel, this attractive term reads $\operatorname{tr}\left(\mathbf{Z}^\top \boldsymbol{L}^\star \mathbf{Z}\right)$ where $\boldsymbol{L}^\star = \mathbb{E}_{\mathbf{W}\sim\mathbb{P}_{\mathcal{P}_X}(\cdot;\boldsymbol{K}_X)}[L(\mathbf{W})]$, boiling down to the objective of Laplacian eigenmaps **belkin2003laplacian**. Therefore, for Gaussian MRFs, the attractive forces resume to an unconstrained Laplacian eigenmaps objective. Such link, already noted in **carreira2010elastic**, is easily unveiled in our framework. Moreover, one can notice that only this attractive term depends on $\mathbf{X}$ as the repulsion is given by the marginal term in (**??**). The latter reads $\mathbb{P}(\mathbf{Z}) = \sum_{\mathbf{W}\in\mathcal{S}_W} \mathbb{P}(\mathbf{Z},\mathbf{W})$ with $\mathbb{P}(\mathbf{Z},\mathbf{W}) \propto f_k(\mathbf{Z},\mathbf{W})\Omega_{\mathcal{P}_Z}(\mathbf{W})$. Such penalty notably prevents a trivial solution, as $\mathbf{0}$, like any constant vector, is a mode of $f_k(\cdot,\mathbf{W})$ for all $\mathbf{W}$. Also note that the prior for $\mathbf{W}_X$ only conditions attraction while the prior for $\mathbf{W}_Z$ only affects repulsion. In the present work we focus solely on deciphering the probabilistic model that accounts for neighbor embedding loss functions and refer to **bohm2020unifying** for a quantitative study of attraction and repulsion in these methods.

**Global Structure Preservation.** To gain intuition, consider that $\mathbf{W}_X$ is observed. As we showed in **??**, when one relies on shift invariant kernels, the positions of the CC means are taken from a diffuse distribution. Since the above methods are all derived from the limit posteriors at $\varepsilon \to 0$, $\mathbf{X}_M$ and $\mathbf{Z}_M$ have no influence on the coupling objective. Hence if two nodes belong to different CCs, their low dimensional pairwise

distance will likely not be faithful. We can expect this phenomenon to persist when the expectation on $\mathbf{W}_X$ is considered, especially when clusters are well distinguishable in $\mathbf{X}$. This observation is central to understand the large scale deficiency of these methods. Note that this happens at the benefit of the local structure which is faithfully represented in low dimension, as discussed in **??**. In the following section we propose to mitigate the global structure deficiency with non-degenerate MRF models.

## 2.4 Towards Capturing Large-Scale Dependencies

In this section, we investigate the ability of graph coupling to faithfully represent global structure in low dimension. To gain intuition on the case where the distribution induced by the graph is not degenerate, we consider a proper Gaussian graph coupling model and show its equivalence with PCA. We then provide a new initialization procedure to alleviate the large scale deficiency of graph coupling when degenerate MRFs are used.

### 2.4.1 PCA as Graph Coupling

As we argue that the inability of SNE-like methods to reproduce the coarse-grain dependencies of the input in the latent space is due to the degeneracy of the conditional (**??**), a natural solution would be to consider graphical models that are well defined and integrable on the entire definition spaces of $\mathbf{X}$ and $\mathbf{Z}$. For simplicity, we consider the Gaussian model and leave the extension to other kernels for future works. Note that in this case integrability translates into the precision matrix being full-rank. As we see with the following, the natural extension of our framework to such models leads to a well-established PCA algorithm. In the following, for a continuous variable $\mathbf{\Theta}_Z$, $\mathbb{P}(\mathbf{\Theta}_Z = \cdot)$ denotes its density.

**Theorem 9.** *Let $\nu \geq n$, $\mathbf{\Theta}_X \sim \mathcal{W}(\nu, \boldsymbol{I}_n)$ and $\mathbf{\Theta}_Z \sim \mathcal{W}(\nu + p - q, \boldsymbol{I}_n)$. Assume that $\mathbf{\Theta}_X$ and $\mathbf{\Theta}_Z$ structure the rows of respectively $\mathbf{X}$ and $\mathbf{Z}$ such that:*

$$\mathrm{vec}(\mathbf{X})|\mathbf{\Theta}_X \sim \mathcal{N}(\mathbf{0}, \mathbf{\Theta}_X^{-1} \otimes \boldsymbol{I}_p), \tag{2.6}$$

$$\mathrm{vec}(\mathbf{Z})|\mathbf{\Theta}_Z \sim \mathcal{N}(\mathbf{0}, \mathbf{\Theta}_Z^{-1} \otimes \boldsymbol{I}_q)\,. \tag{2.7}$$

*Then the solution of the precision coupling problem:*

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times q}} -\mathbb{E}_{\mathbf{\Theta}_X|\mathbf{X}} \left[\log \mathbb{P}(\mathbf{\Theta}_Z = \mathbf{\Theta}_X|\mathbf{Z})\right]$$

*is a PCA embedding of $\mathbf{X}$ with $q$ components.*

We now highlight the parallels with the previous construction done for neighbor embedding methods. First note that the multivariate Gaussian with full-rank precision is inherently a pairwise MRF **rue2005gaussian**. When choosing the Gaussian kernel for neighbor embedding methods, we saw that the graph Laplacian $\boldsymbol{L}_X$ of $\mathbf{W}_X$ was playing the role of the among-row precision matrix, as we had $\mathbf{X}|\mathbf{W}_X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{L}_X^{-1} \otimes \boldsymbol{I}_p)$ (equation **??**). Recall that the later always has a null-space which is spanned by the CC indicator vectors of $\mathbf{W}$ (**??**). Here, the key difference is that we impose a full-rank constraint on the precision $\mathbf{\Theta}$. Concerning the priors, we choose the ones that are conjugate to the conditionals (**??**) and (**??**), as previously done when constructing the prior for neighbor embedding methods (definition **??**). Hence in the full-rank setting, the prior simply amounts to a Wishart distribution denoted by $\mathcal{W}$.

The above theorem further highlights the flexibility and generality of the graph coupling framework. Unlike usual constructions of PCA or probabilistic PCA **tipping1999probabilistic**, in the above the linear relation between $\mathbf{X}$ and $\mathbf{Z}$ is recovered by solving the graph coupling problem and not explicitly stated beforehand. To the best of our knowledge, it is the first time such a link is uncovered between PCA and SNE-like methods. In contrast with the latter, PCA is well-known for its ability to preserve global structure while being significantly less efficient at identifying clusters **anowar2021conceptual**. Therefore, as suspected in **??**, the degeneracy of the conditional distribution given the graph is key to determine the distance preservation properties of the embeddings. We propose in **??** to combine both graph coupling approaches to strike a balance between global and local structure preservation.

### 2.4.2 Hierarchical Graph Coupling

The goal of this section is to show that global structure in SNE-like embeddings can be improved by structuring the CCs' positions. We consider the following hierarchical model for $\mathbf{X}$, where $\mathcal{P}_X \in \{B, D, E\}$, $k_x$ satisfies the assumptions of **??** and $\nu_X \geq n$:

$$\begin{aligned} \mathbf{W}_X \sim \mathbb{P}^{\varepsilon}_{\mathcal{P}_x, k_x}(\cdot\,; \mathbf{1}, 1), \quad \boldsymbol{\Theta}_X | \mathbf{W}_X \sim \mathcal{W}(\nu_X, \boldsymbol{I}_R) \\ \mathbf{X}_C | \mathbf{W}_X \sim \mathbb{P}_{k_x}(\cdot\,|\mathbf{W}_X), \quad \operatorname{vec}(\mathbf{X}_M) | \boldsymbol{\Theta}_X \sim \mathcal{N}\left(\mathbf{0}, \left(\varepsilon \boldsymbol{U}_{[:R]} \boldsymbol{\Theta}_X \boldsymbol{U}_{[R]}^\top\right)^{-1} \otimes \boldsymbol{I}_p\right) \end{aligned}$$

where $\boldsymbol{U}_{[R]}$ are the eigenvectors associated to the Laplacian null-space of $\overline{\mathbf{W}}_X$. Given a graph $\mathbf{W}_X$, the idea is to structure the CCs' relative positions with a full-rank Gaussian model. The same model is considered for $\mathbf{W}_Z$, $\boldsymbol{\Theta}_Z$ and $\mathbf{Z}$, choosing $\nu_Z = \nu_X + p - q$ for the Wishart prior to satisfy the assumption of **??**. With this in place, we aim at providing a complete coupling objective, matching the pairs $(\mathbf{W}_X, \boldsymbol{\Theta}_X)$ and $(\mathbf{W}_Z, \boldsymbol{\Theta}_Z)$. The joint negative cross-entropy can be decomposed as follows:

$$\begin{aligned} &\mathbb{E}_{(\mathbf{W}_X, \boldsymbol{\Theta}_X)|\mathbf{X}} \left[\log \mathbb{P}((\mathbf{W}_Z, \boldsymbol{\Theta}_Z) = (\mathbf{W}_X, \boldsymbol{\Theta}_X) | \mathbf{Z})\right] \\ &= \mathbb{E}_{\mathbf{W}_X|\mathbf{X}} \left[\log \mathbb{P}(\mathbf{W}_Z = \mathbf{W}_X | \mathbf{Z})\right] + \qquad (2.8) \\ &\mathbb{E}_{(\mathbf{W}_X, \boldsymbol{\Theta}_X)|\mathbf{X}} \left[\log \mathbb{P}(\boldsymbol{\Theta}_Z = \boldsymbol{\Theta}_X | \mathbf{W}_Z = \mathbf{W}_X, \mathbf{Z})\right] \qquad (2.9) \end{aligned}$$

where (**??**) is the usual coupling criterion of $\mathbf{W}_X$ and $\mathbf{W}_Z$ capturing intra-CC variability while (**??**) is a penalty resulting from the Gaussian structure on $\mathcal{S}_M$. Constructed as such, the above objective allows a trade-of between local and global structure preservation. Following current trends in DR **kobak2021initialization**, we propose to take care of the global structure first *i.e.* focusing on (**??**) before (**??**). The difficulty of dealing with (**??**) lies in the hierarchical construction of the graph and the Gaussian precision (see **??**). We state the following result.

**Corollary 10.** *Let $\mathbf{W}_X \in \mathcal{S}_W$, $\boldsymbol{L} = L(\overline{\mathbf{W}}_X)$ and $\mathcal{S}_M^q = (\ker \boldsymbol{L}) \otimes \mathbb{R}^q$, then for all $\varepsilon > 0$, given the above hierarchical model, the solution of the problem:*

$$\min_{\mathbf{Z} \in \mathcal{S}_M^q} -\mathbb{E}_{\boldsymbol{\Theta}_X|\mathbf{X}} \left[\log \mathbb{P}(\boldsymbol{\Theta}_Z = \boldsymbol{\Theta}_X | \mathbf{W}_Z = \mathbf{W}_X, \mathbf{Z})\right]$$

*is a PCA embedding of $\boldsymbol{U}_{[:R]} \boldsymbol{U}_{[R]}^\top \mathbf{X}$ where $\boldsymbol{U}_{[:R]}$ are the CCs' membership vectors of $\overline{\mathbf{W}}_X$.*

**Remark 11.** Note that while (**??**) approximates the objective of SNE-like methods when $\varepsilon \to 0$, the minimizer of (**??**) given by **??** is stable for all $\varepsilon$.

From this observation, we propose a simple heuristic to minimize (**??**) that consists in computing a PCA embedding of $\mathbb{E}_{\mathbb{P}_{\mathcal{P}_X}(\cdot;\boldsymbol{K}_X)}\left[\boldsymbol{U}_{[:R]}\boldsymbol{U}_{[R]}^\top\right]\mathbf{X}$. The distribution of the connected components of the posterior of $\mathbf{W}_X$ being intractable, we resort to a Monte-Carlo estimation of the above expectation. The latter procedure called *ccPCA* aims at recovering the inter-CC structure that is filtered by SNE-like methods. *ccPCA* may then be used as initialization for optimizing (**??**) which is done by running the DR method corresponding to the graph priors at hand (**??**). This second step essentially consists in refining the intra-CC structure.

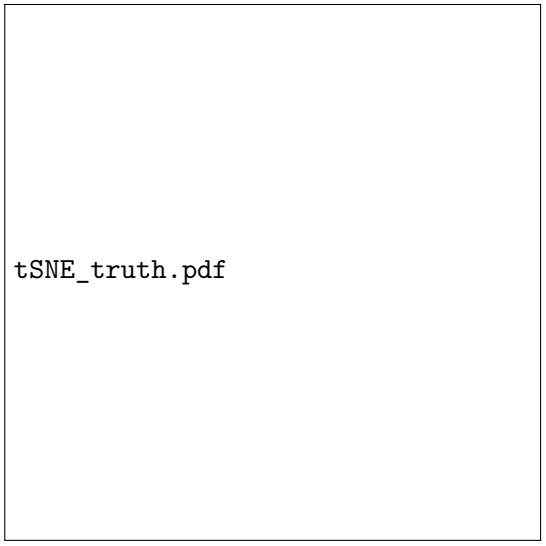


FIGURE 2.1: Left: MNIST t-SNE (perp : 30) embeddings initialized with i.i.d $\mathcal{N}(0,1)$ coordinates. Middle: using these t-SNE embeddings, mean coordinates for each digit are represented. Right: we compute a matrix of mean input coordinates for each of the 10 digits and embed it using PCA. Digit clusters in the graph positions of clusters vary across different **mnaaten2008tSNE** runs but visually match the PCA embeddings of input mean vectors (right plot).

### 2.4.3 Experiments with *ccPCA*

**??** shows that a t-SNE embedding of a balanced MNIST dataset of 10000 samples **deng2012mnist** with isotropic Gaussian initialization performs poorly in conserving the relative positions of clusters. As each digit cluster contains approximately 1000 points, with a perplexity of 30, sampling an edge according to the graph posterior $\mathbb{P}_{\mathcal{P}_X}(\cdot;\boldsymbol{K}_X)$ is very unlikely. Recall that the perplexity value corresponds to the approximate number of effective neighbors of each point. Hence images of different digits are with very high probability in different CCs of the graph posterior and their CC-wise means are not coupled as discussed in **??**. To remedy this in practice, PCA or Laplacian eigenmaps are usually used as initialization **kobak2021initialization**.

These strategies are tested (**??**) together with *ccPCA*. This shows that *ccPCA* manages to retrieve the digits that mostly support the large-scale variability as measured by the peripheral positioning of digits 0 (blue), 2 (green), 6 (pink) and 7 (grey) given by the right side of **??**. Other perplexity values for *ccPCA* are explored in appendix **??** while the experimental setup is detailed in appendix **??**. In appendix **??**, we perform quantitative evaluations of *ccPCA* for both t-SNE and UMAP on various datasets using K-ary neighborhood criteria. We find that using *ccPCA* as initialization is in general more reliable than PCA and Laplacian eigenmaps for preserving global structure using both t-SNE and UMAP.

Compared to PCA, *ccPCA* manages to aggregate points into clusters, thus filtering the intra-cluster variablity and focusing solely on the inter-cluster structure. Compared to Laplacian eigenmaps which performs well at identifying clusters but suffers from the same deficiency as t-SNE for positioning them, *ccPCA* retains more of the coarse-grain structure. These observations support our unifying probabilistic framework and the theoretical results about the MRF degeneracy which are the leading contributions of this article. The *ccPCA* initialization appears as a first stepping stone towards more grounded DR methods based on the probabilistic model presented in this article.

## 2.5 Conclusion and Perspectives

In this work we shed a new light on most popular DR methods by showing that they can be unified within a common probabilistic model in the form of latent Markov

Random Fields Graphs coupled by a cross entropy. The definition of such a model constitutes a major step towards the understanding of common dimension reduction methods, in particular their structure preservation properties as discussed in this article.

Our work offers many perspectives, among which the possibility to enrich the probabilistic model with more suited graph priors. Currently considered priors are simply the ones that are conjugate to the MRFs thus they are mostly designed to yield a tractable coupling objective. However they may not be optimal and could be modified to capture targeted features, *e.g.* communities, in the input data, and give adapted representations in the latent space. The graph coupling approach could also be extended to more general latent structures governing the joint distribution of observations. Finally, the probabilistic model could be leveraged to tackle hyper-parameter calibration, especially kernel bandwidths that have a great influence on the quality of the representations and are currently tuned using heuristics with unclear motivations.

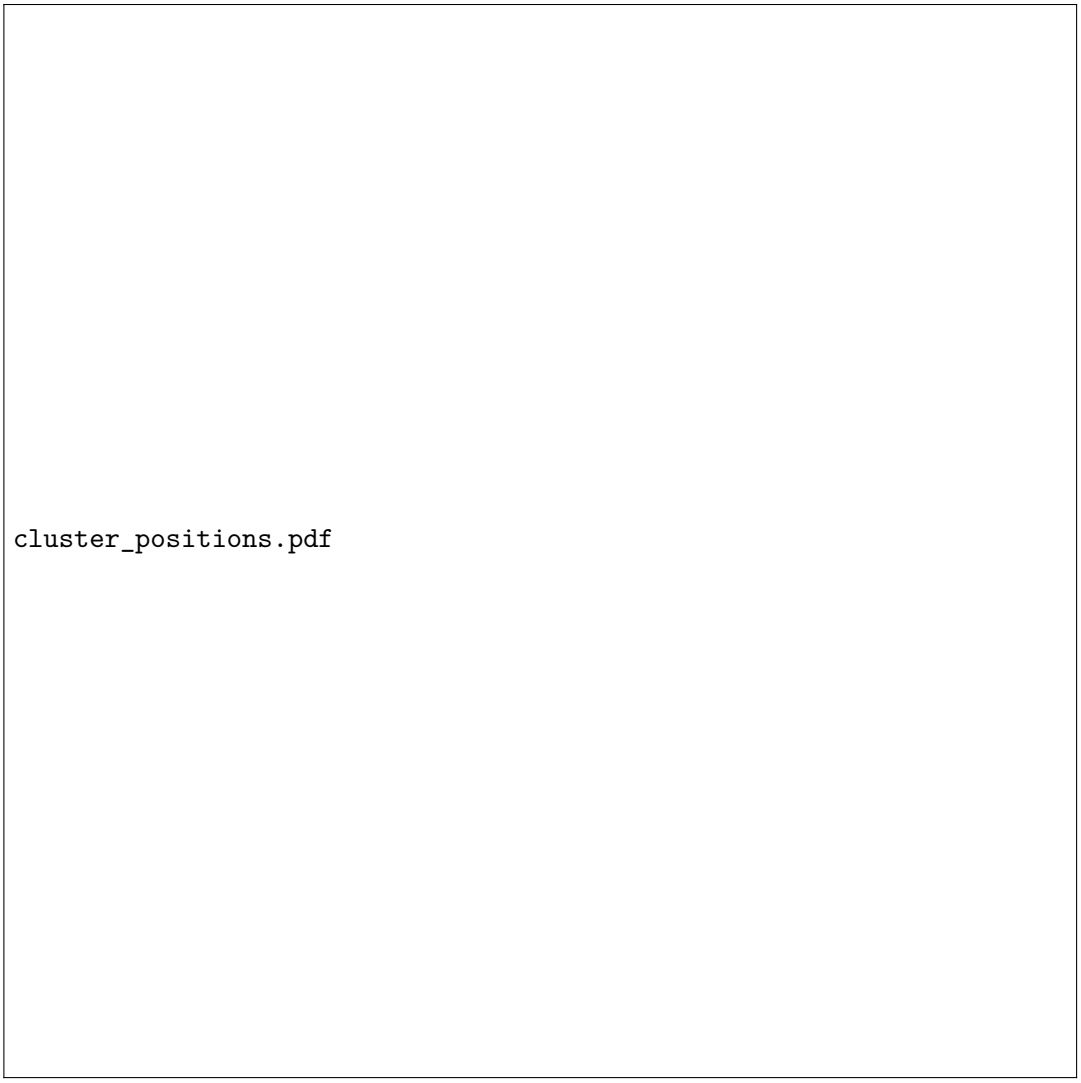


FIGURE 2.2: Top: MNIST embeddings produced by PCA, Laplacian eigenmaps, *ccPCA* and finally t-SNE launched after the previous three embeddings to improve the fine-grain structure. Bottom: mean coordinates for each digit using the embeddings of the first row. The color legend is the same as in **??**. t-SNE was trained during 1000 iterations using default parameters with the openTSNE implementation **polivcar2019opentsne**.

# Appendix A

# Frequently Asked Questions

## A.1 How do I change the colors of links?

The color of links can be changed to your liking using:

`\hypersetup{urlcolor=red}`, or

`\hypersetup{citecolor=green}`, or

`\hypersetup{allcolor=blue}`.

If you want to completely hide the links, you can use:

`\hypersetup{allcolors=.}`, or even better:

`\hypersetup{hidelinks}`.

If you want to have obvious links in the PDF but not the printed text, use:

`\hypersetup{colorlinks=false}`.