2D Visualization of Immune System Cellular Protein Data by Nonlinear Dimensionality Reduction

Andre Esteva ESTEVA@STANFORD.EDU
Stanford University, Electrical Engineering, 496 Lomita Mall, Durand 196, Stanford, CA 94305 USA

Anand Sampat ASAMPAT@STANFORD.EDU
Stanford University, Electrical Engineering, 450 Serra Mall, Stanford, CA 94305 USA

Amit Badlani ABADLANI@STANFORD.EDU
Stanford University, Electrical Engineering, 450 Serra Mall, Stanford, CA 94305 USA

Abstract

We present in this paper a way to effectively visualize multi-dimensional immune system cellular data by means of nonlinear methods. We find that Stochastic Neighbor Embedding (SNE), and it's variations, t-SNE and s-SNE, to be most effective at successfully mapping clusters of points into a two dimensional embedding space while preserving both the structure between similar points and the disparity between different clusters. Using a centroid-based metric that relabels points according to the cluster centroid to which they are closest, we conclude that SNE works significantly better than linear and spectral methods. In addtional, by using an optimization approach for SNE similar to Newton's Method, but with the Hessian of the objective function, $\nabla^2(E)$, replaced by its first term, we are able to run the SNE varients and EE two orders of magnitude faster than with standard optimization.

1. Introduction

1.1. Immune Cell Data

In the field of cancer immunology, scientists use the protein content of immune system cells as a way to identify a cells corresponding type. For example, immune system cells, which are contained in bone marrow, are comprised of a variety of cell types, and to a large degree are uniquely identifiable by the proteins they contain. Highly sophisticated methods have been developed that process cells and return information on the types and quantities of proteins expressed in those cells. This data can then be viewed by an expert in the field and categorized. The laborious process of viewing the different dimensions of protein expression and categorizing a cell's type is known as gating. Figure 1.1, below, taken from (Amir et al., 2013) shows this graphically.

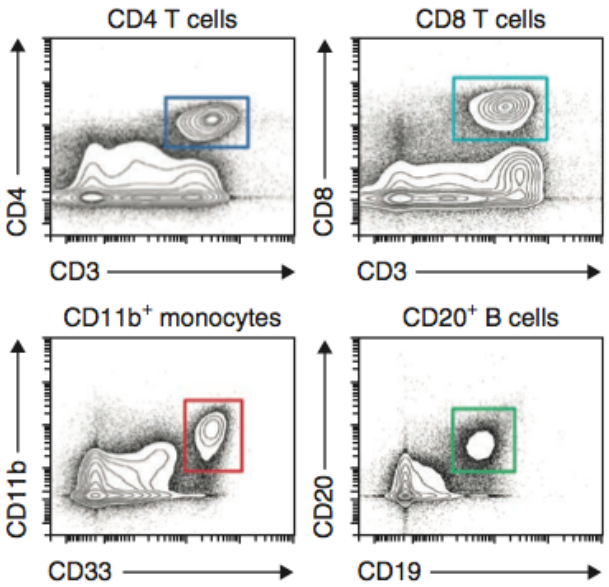

Figure 1. Strategy for cell gating: Two single dimensions of a cell are viewed at a time and through an iterative process the cell is classified

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

1.2. Project Goals and Metrics

It is of interest to cancer immunologists to find structure within multi-dimensional protein expression space and map it onto a lower dimensions (refered to henceforth as a map space) for ease of visualization and understanding. As cells change and evolve, so too do the types and quantities of the proteins they express. This leads to a shifting of their representation in multi-dimensional space which can be tracked. Dimensionality reduction of original biological data coupled with a metric for how well the projection represents the original data would provide biologists with a powerful tool for understanding the structure of their data. To address these challenges we demonstrate:

- The application of linear and non-linear methods of dimensionality reduction of multi-dimensional protein data
- A metric-based comparison of how each algorithm performs

2. Data Representation

2.1. Data Acquisition

Mass cytometry is a single-cell multiparametric protein detection technology based on inductively coupled plasma mass spectrometry. It is an extension of flow cytometry in which antibodies are tagged with isotopically pure rare earth elements allowing simultaneous measurement of greater than 40 parameters while circumventing the issue of spectral overlap. In single-cell droplet form, the cells are passed through an elemental mass spectrometer and an integrator to generate an $m \times p$ matrix where m is the number of cells processed and p is the number of distinct proteins contained in the cell set. These matrices are stored as .FCS files in online databanks which we have been granted access to. Figure 2.1 shows this process.

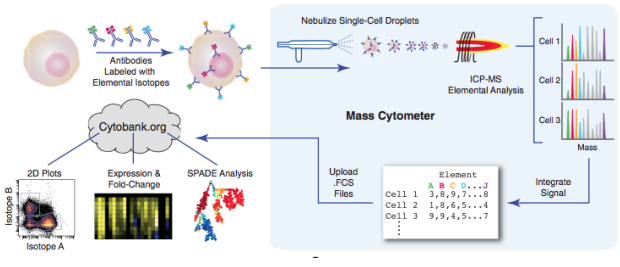

Figure 2. Overview of data acquisition, from extraction of cellular protein counts to storage in online databanks

Table 1. Classification of Cell Types and their Corresponding Sub-Types

Cell Types	Sub-types
Stem Cell	HSC, MPP, CMP, GMP, MEP
B Cells	Plasma, pre-B-I, pre-B- II, Immature, Mature CD38 low, Mature CD38 mid
T Cells	Mature CD4+, Mature CD8+ Naive CD4+, Naive CD8+
NK	-
pDC	-
Monocytes	CD11B - , CD11B high, CD11B mid

Table 1 lists the cell types and subtypes that are parsed using this method.

2.2. Feature Selection

The $p = 41$ protein counts collected for each cell that passes through mass cytometry is comprised of both intracellular and surface proteins. These two types play fundamentally different roles in cell identification. Surface proteins are semi-permanent markers that last for significant periods of time relative to the lifetime of a cell, whereas intracellular proteins are highly transient and can change quickly. This is analogous to classifying a person based on where they live (semi-permanent) versus what they wore on a particular day (transient). Understanding this, we select as our feature space the $n = 17$ surface protein markers of the cell data.

3. Methods

The datafiles provided contain cell counts on the order of tens of thousands where we consider each cell to be a point in $\mathbb{R}^n$. To simplify our algorithms and account for matrix size differences in the difference .FCS files we run our algorithms on equally sized portions of different cell data. In particular, if we let $\mathbb{S}$ be the set of all cell sub-types as defined in Table 1, $S \in \mathbb{S}$ be some subset of interest with cardinality $|S|$, and N some fixed positive integer, then by taking N rows from each sub-type $s \in S$ we form a matrix $M \in \mathbb{R}^{N|S| \times n}$ on which we can run algorithms quickly and without giving unfair weighting to a particular cell sub-type.

We consider various algorithms which project sets of data in $\mathbb{R}^n$ onto $\mathbb{R}^2$ for easy visualization.

3.1. Linear Methods

3.2. Spectral Methods

3.3. Nonlinear Methods

4. Electronic Submission

As in the past few years, ICML will rely exclusively on electronic formats for submission and review.

4.1. Templates for Papers

Electronic templates for producing papers for submission are available for LaTeX and Microsoft Word. Templates are accessible on the World Wide Web at: `http://icml.cc/2012/`

Send questions about these electronic templates to `program@icml.cc`.

The formatting instructions below will be enforced for initial submissions and camera-ready copies.

- The maximum paper length is 8 pages.
- Do not alter the style template; in particular, do not compress the paper format by reducing the vertical spaces.
- Do not include author information or acknowledgments in your initial submission.
- Place figure captions *under* the figure (and omit titles from inside the graphic file itself). Place table captions *over* the table.
- References must include page numbers whenever possible and be as complete as possible. Place multiple citations in chronological order.

Please see below for details on each of these items.

4.2. Submitting Papers

Submission to ICML 2012 will be entirely electronic, via a web site (not email). The URL and information about the submission process are available on the conference web site at

`http://icml.cc/2012/`

Paper Deadline: The deadline for paper submission to ICML 2012 is Friday, February 24, 2012, at 23:59 Universal Time (3:59 Pacific Daylight Time). If your full submission does not reach us by this date, it will not be considered for publication. There is no separate abstract submission.

Anonymous Submission: To facilitate blind review, no identifying author information should appear on the title page or in the paper itself. Section 5.3 will explain the details of how to format this.

Simultaneous Submission: ICML will not accept any paper which, at the time of submission, is under review for another conference or has already been published. This policy also applies to papers that overlap substantially in technical content with conference papers under review or previously published. ICML submissions must not be submitted to other conferences during ICML's review period. Authors may submit to ICML substantially different versions of journal papers that are currently under review by the journal, but not yet accepted at the time of submission. Informal publications, such as technical reports or papers in workshop proceedings which do not appear in print, do not fall under these restrictions.

To ensure our ability to print submissions, authors must provide their manuscripts in **PDF** format. Furthermore, please make sure that files contain only Type-1 fonts (e.g., using the program `pdffonts` in linux or using File/DocumentProperties/Fonts in Acrobat). Other fonts (like Type-3) might come from graphics files imported into the document.

Authors using **Word** must convert their document to PDF. Most of the latest versions of Word have the facility to do this automatically. Submissions will not be accepted in Word format or any format other than PDF. Really. We're not joking. Don't send Word.

Those who use **LaTeX** to format their accepted papers need to pay close attention to the typefaces used. Specifically, when producing the PDF by first converting the dvi output of LaTeX to Postscript the default behavior is to use non-scalable Type-3 PostScript bitmap fonts to represent the standard LaTeX fonts. The resulting document is difficult to read in electronic form; the type appears fuzzy. To avoid this problem, dvips must be instructed to use an alternative font map. This can be achieved with something like the following commands:

dvips -Ppdf -tletter -G0 -o paper.ps paper.dvi
ps2pdf paper.ps

Note that it is a zero following the "-G". This tells dvips to use the config.pdf file (and this file refers to a better font mapping).

Another alternative is to use the **pdflatex** program instead of straight LaTeX. This program avoids the Type-3 font problem, however you must ensure that all of the fonts are embedded (use `pdffonts`). If they are not, you need to configure pdflatex to use a font map file that specifies that the fonts be embedded. Also

you should ensure that images are not downsampled or otherwise compressed in a lossy way.

Note that the 2012 style files use the `hyperref` package to make clickable links in documents. If this causes problems for you, add `nohyperref` as one of the options to the `icml2012` usepackage statement.

4.3. Reacting to Reviews

We will continue the ICML tradition in which the authors are given the option of providing a short reaction to the initial reviews. These reactions will be taken into account in the discussion among the reviewers and area chairs.

4.4. Submitting Final Camera-Ready Copy

The final versions of papers accepted for publication should follow the same format and naming convention as initial submissions, except of course that the normal author information (names and affiliations) should be given. See Section 5.3.2 for details of how to format this.

The footnote, "Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute." must be modified to "Appearing in *Proceedings of the 29^{th} International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012. Copyright 2012 by the author(s)/owner(s)."

For those using the LaTeX style file, simply change `\usepackage{icml2012}` to

```
\usepackage[accepted]{icml2012}
```

Authors using **Word** must edit the footnote on the first page of the document themselves.

Camera-ready copies should have the title of the paper as running head on each page except the first one. The running title consists of a single line centered above a horizontal rule which is 1 point thick. The running head should be centered, bold and in 9 point type. The rule should be 10 points above the main text. For those using the LaTeX style file, the original title is automatically set as running head using the `fancyhdr` package which is included in the ICML 2012 style file package. In case that the original title exceeds the size restrictions, a shorter form can be supplied by using

```
\icmltitlerunning{...}
```

just before `\begin{document}`. Authors using **Word** must edit the header of the document themselves.

5. Format of the Paper

All submissions must follow the same format to ensure the printer can reproduce them without problems and to let readers more easily find the information that they desire.

5.1. Length and Dimensions

Papers must not exceed eight (8) pages, including all figures, tables, references, and appendices. Any submission that exceeds this page limit or that diverges significantly from the format specified herein will be rejected without review.

The text of the paper should be formatted in two columns, with an overall width of 6.75 inches, height of 9.0 inches, and 0.25 inches between the columns. The left margin should be 0.75 inches and the top margin 1.0 inch (2.54 cm). The right and bottom margins will depend on whether you print on US letter or A4 paper, but all final versions must be produced for US letter size.

The paper body should be set in 10 point type with a vertical spacing of 11 points. Please use Times Roman typeface throughout the text.

5.2. Title

The paper title should be set in 14 point bold type and centered between two horizontal rules that are 1 point thick, with 1.0 inch between the top rule and the top edge of the page. Capitalize the first letter of content words and put the rest of the title in lower case.

5.3. Author Information for Submission

To facilitate blind review, author information must not appear. If you are using LaTeX and the `icml2012.sty` file, you may use `\icmlauthor{...}` to specify authors. The author information will simply not be printed until `accepted` is an argument to the style file. Submissions that include the author information will not be reviewed.

5.3.1. Self-Citations

If your are citing published papers for which you are an author, refer to yourself in the third person. In particular, do not use phrases that reveal your identity (e.g., "in previous work (Bendall et al., 2011), we have shown ...").

Do not anonymize citations in the reference section by removing or blacking out author names. The only exception are manuscripts that are not yet published

(e.g. under submission). If you choose to refer to such unpublished manuscripts (**?**), anonymized copies have to be submitted as Supplementary Material via CMT. However, keep in mind that an ICML paper should be self contained and should contain sufficient detail for the reviewers to evaluate the work. In particular, reviewers are not required to look a the Supplementary Material when writing their review.

5.3.2. Camera-Ready Author Information

If a paper is accepted, a final camera-ready copy must be prepared. For camera-ready papers, author information should start 0.3 inches below the bottom rule surrounding the title. The authors' names should appear in 10 point bold type, electronic mail addresses in 10 point small capitals, and physical addresses in ordinary 10 point type. Each author's name should be flush left, whereas the email address should be flush right on the same line. The author's physical address should appear flush left on the ensuing line, on a single line if possible. If successive authors have the same affiliation, then give their physical address only once.

A sample file (in PDF) with author names is included in the ICML2012 style file package.

5.4. Abstract

The paper abstract should begin in the left column, 0.4 inches below the final address. The heading 'Abstract' should be centered, bold, and in 11 point type. The abstract body should use 10 point type, with a vertical spacing of 11 points, and should be indented 0.25 inches more than normal on left-hand and right-hand margins. Insert 0.4 inches of blank space after the body. Keep your abstract brief and self-contained, limiting it to one paragraph and no more than six or seven sentences.

5.5. Partitioning the Text

You should organize your paper into sections and paragraphs to help readers place a structure on the material and understand its contributions.

5.5.1. Sections and Subsections

Section headings should be numbered, flush left, and set in 11 pt bold type with the content words capitalized. Leave 0.25 inches of space before the heading and 0.15 inches after the heading.

Similarly, subsection headings should be numbered, flush left, and set in 10 pt bold type with the content words capitalized. Leave 0.2 inches of space before the heading and 0.13 inches afterward.

Finally, subsubsection headings should be numbered, flush left, and set in 10 pt small caps with the content words capitalized. Leave 0.18 inches of space before the heading and 0.1 inches after the heading.

Please use no more than three levels of headings.

5.5.2. Paragraphs and Footnotes

Within each section or subsection, you should further partition the paper into paragraphs. Do not indent the first line of a given paragraph, but insert a blank line between succeeding ones.

You can use footnotes[1] to provide readers with additional information about a topic without interrupting the flow of the paper. Indicate footnotes with a number in the text where the point is most relevant. Place the footnote in 9 point type at the bottom of the column in which it appears. Precede the first footnote in a column with a horizontal rule of 0.8 inches.[2]

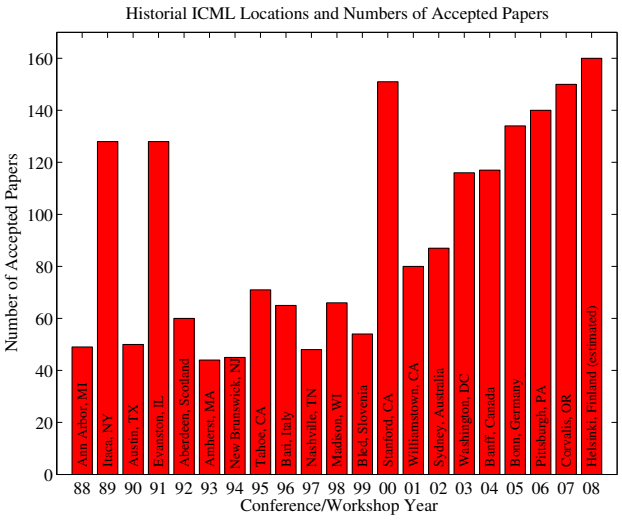

Figure 3. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

[1]For the sake of readability, footnotes should be complete sentences.

[2]Multiple footnotes can appear in each column, in the same order as they appear in the text, but spread them across columns and pages if possible.

Algorithm 1 Bubble Sort

```
Input: data x_i, size m
repeat
  Initialize noChange = true.
  for i = 1 to m − 1 do
    if x_i > x_{i+1} then
      Swap x_i and x_{i+1}
      noChange = false
    end if
  end for
until noChange is true
```

Table 2. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

Data set	Naive	Flexible	Better?
Breast	95.9± 0.2	96.7± 0.2	√
Cleveland	83.3± 0.6	80.0± 0.6	×
Glass2	61.9± 1.4	83.8± 0.7	√
Credit	74.8± 0.5	78.3± 0.6	
Horse	73.3± 0.9	69.7± 1.0	×
Meta	67.1± 0.6	76.5± 0.5	√
Pima	75.1± 0.6	73.9± 0.5	
Vehicle	44.9± 0.6	61.5± 0.4	√

5.6. Figures

You may want to include figures in the paper to help readers visualize your approach and your results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should serve this function.

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space before the caption and 0.1 inches after it, as in Figure 3. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX), but always place two-column figures at the top or bottom of the page.

5.7. Algorithms

If you are using LaTeX, please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

5.8. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the table with at least 0.1 inches of space before the title and the same after it, as in Table 2. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

Tables contain textual material that can be typeset, as contrasted with figures, which contain graphical material that must be drawn. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns, but place two-column tables at the top or bottom of the page.

5.9. Citations and References

Please use APA reference format regardless of your formatter or word processor. If you rely on the LaTeX bibliographic facility, use `natbib.sty` and `icml2012.bst` included in the style-file package to obtain this format.

Citations within the text should include the authors' last names and year. If the authors' names are included in the sentence, place only the year in parentheses, for example when referencing Arthur Samuel's pioneering work (2011). Otherwise place the entire reference in parentheses with the authors and year separated by a comma (Bendall et al., 2011). List multiple references separated by semicolons (Bendall et al., 2011; Crammer & Singer, 2002; Bendall et al., 2011). Use the 'et al.' construct only for citations with three or more authors or after listing all authors to a publication in an earlier reference (Bendall et al., 2011).

Authors should cite their own work in the third person in the initial version of their paper submitted for blind review. Please refer to Section 5.3 for detailed instructions on how to cite your own papers.

Use an unnumbered first-level section heading for the references, and use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 10 points. The references at the end of this document give examples for journal articles (Bendall et al., 2011), conference publications (Bendall et al., 2011), book chapters (Bendall

et al., 2011), books (Bendall et al., 2011), edited volumes (Bendall et al., 2011), technical reports (Crammer & Singer, 2002), and dissertations (Bendall et al., 2011).

Alphabetize references by the surnames of the first authors, with single author entries preceding multiple author entries. Order references for the same authors by year of publication, with the earliest first. Make sure that each reference includes all relevant information (e.g., page numbers).

5.10. Software and Data

We strongly encourage the publication of software and data with the camera-ready version of the paper whenever appropriate. This can be done by including a URL in the camera-ready copy. However, do not include URLs that reveal your institution or identity in your submission for review. Instead, provide an anonymous URL or upload the material as "Supplementary Material" into the CMT reviewing system. Note that reviewers are not required to look a this material when writing their review.

Acknowledgments

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and probably should) include acknowledgements. In this case, please place such acknowledgements in an unnumbered section at the end of the paper. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

References

Amir, El-ad David, Davis, Kara L, Tadmor, Michelle D, Simonds, Erin F, Levine, Jacob H, Bendall, Sean C, Shenfeld, Daniel K, Krishnaswamy, Smita, Nolan, Garry P, and Pe'er, Dana. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, 31(6):545–552, June 2013.

Bendall, S C, Simonds, E F, Qiu, P, Amir, E a D, Krutzik, P O, Finck, R, Bruggner, R V, Melamed, R, Trejo, A, Ornatsky, O I, Balderas, R S, Plevritis, S K, Sachs, K, Pe'er, D, Tanner, S D, and Nolan, G P. Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science*, 332(6030):687–696, May 2011.

Crammer, Koby and Singer, Yoram. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292, March 2002.