Identifying Feature, Parameter, and Sample Subsets in Machine Learning and Image Analysis

by

Ronak Mehta

A preliminary report submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Computer Sciences Department)

at the

UNIVERSITY OF WISCONSIN–MADISON

2022

Date of preliminary examination: 2022-05-12

The examination is approved by the following members of the Oral Committee:

Frederic Sala, Assistant Professor, Computer Sciences Department
Yong Jae Lee, Associate Professor, Computer Sciences Department
Michael Newton, Professor, Statistics Deparment
Vikas Singh (Advisor), Professor, Biostatistics and Medical Informatics Department

Contents

Chapter 1

Introduction

Modern applications of machine learning have become ubiquitous, to the point where almost all interactions with so-called “smart" technologies involve some call and response with some form of learned predictor. The large scale of these models, up to billions of parameters in neural networks, makes possible what would otherwise be an unconstrained, infeasible learning task via a highly parameterized model. These parameterizations have enabled human-level performance on learning tasks previously thought to have been insurmountable for any potential artificial intelligence system. The high accuracy and generalization of these large scale models notwithstanding, a number of parallel questions have arisen and are gaining prominence with respect to the performance of these models. What features are most important for prediction? Which samples were most important for my training? Can we understand when a model is certain or uncertain about its output? Are there layers in my network that have learned a particular subtask? Questions of robustness, bias, influence, fairness, and importance have become central questions to contemporary machine learning research (Doshi-Velez and Kim, 2017; Mehrabi et al., 2021; Amodei et al., 2016).

Traditional statistical learning methods have been studied with these questions in mind for many decades, and have found new life in these subfields. Linear regressors, decision trees, and support vector machines have all been analyzed under these lenses, and as the modern machine learning community has returned to these questions so has a renewed interest in their methods of analysis. New research focuses particularly on the differences associated with moving from classical under-parametrized models to modern **over-parameterized** models: where the model dimension vastly outnumbers the number of input samples, and may even be comparable to the *entire*

sample space. While nascent, these new methods and analyses attempt to fill the gap between statistical and deep models to enable similar measures of sample influence, feature importance, and model understanding.

A full picture. Consider a dataset $\{X\}_{i=1}^{n}$ of size n where each data point in the set X is considered drawn from some underlying distribution over the domain $X \sim \mathcal{X}^{d}$, with domain dimensionality d. A model f is fit using a parametrizations $\theta \in \Theta$, with Θ the space of possible parametrization with some intrinsic dimension p. From an analysis perspective, we might be interested in any one of (a) subsets of input features $\mathcal{C} \subseteq \mathcal{X}$ that are important for the downstream task, (b) associating model subsets $\mathcal{P} \subseteq \Theta$ with specific inputs or groups of inputs, or (c) subsets or subgroups of samples $S \subseteq \{X\}^{n}$ equally of interest. While all three of these problems are closely related, they require different approaches.

This thesis *will focus its main efforts on identifying these important subsets of model, feature, and sample space for feature association, model size reduction, model unlearning, and, fairness. Specifically, taking advantage of both existing statistical and geometric methods, we will develop new methods for localizing subsets, in a range of settings from hypothesis testing on the one hand to deep learning frameworks on the other.*

Feature selection in the case of typical regression or classification takes some form of learning parameters θ that allow for $\hat{y} = f_{\theta}(x)$ to be close to the true outcome of interest y. While forms of data $X := (x, y)$ may simply be continuous and real-valued, modern machine learning has greatly expanded formulations of the classical learning problem to include a wide variety of structured learning problems (Nowozin and Lampert, 2011). Consider the case when a high-dimensional input is used to predict an output with a highly-parametrized model. Once learned, obvious questions arise as discussed above: are there specific low-dimensional spaces in either the input or the model space that are most important or necessary for the global learning problem of interest? Are there specific subspaces associated with particular subproblems of the global problem? The machine learning literature has come up with a number of ways to identify analogs of these spaces, including extensions of sensitivity analysis to deep learning (Yeung et al., 2010; Zhang and Wallace, 2015), and constructing and identifying nonzero model subsets via particular model choices such as activations (Selvaraju et al., 2017) and regularizers. In classical settings these are well understood:

decision trees naturally provide ease of interpretibility via the information used to choose splits, and both linear and kernel support vector machines have been analyzed to provide for measures of sample importance via distances to the margin as well as feature importance via weights defining the learned hyperplane (Mitchell, 1997). Attention and saliency maps have emerged as popular new methods, given their ease of implementation and interpretation (Sutskever et al., 2014; Vaswani et al., 2017; Selvaraju et al., 2017). By learning dimensions of a given input that are particularly important, either in a hard (binary) or soft (continuous weighting) manner, model builders are better able to understand and interpret what a model has learnt.

The specific ideas of attention notwithstanding, many of these existing methods are far removed from traditional hypothesis testing frameworks., While some work has begun in this direction (Tansey et al., 2018), there remains a gap in direct identification of subsets and structures in these spaces that can be defined in statistically rigorous manners.

A specific example. Consider a traditional machine learning classification task in which we would like to predict whether an individual has a specific disease condition based on a medical resonance image (MRI) scan of their brain. Our input feature x may consist of a 3D-array of values lying in $\mathbb{R}^{x \times y \times z}$ measuring some intensity of the imaging modality at each voxel, indexed by a tuple (i, j, k). Our outcome variable y may simply be a binary label of whether the input scan has been labeled by a radiologist as one demonstrating typical disease characteristics. Using an off the shelf 3D convolutional neural network with adjustments to match our input size, we can very quickly set up and train a system to predict disease presence with a high degree of accuracy.

While attention can be directly applied to the network in order to identify “hotspots" in the input space relating to the learned classification task, given the high-dimensional nature of the input and the relatively small sample size associated with medical imaging data, it is very likely that an area of interest identified may be an intricacy of the training samples used rather than truly a region of disease signal. Class activation maps (CAMs) may be unclear, and can often associate with image artifacts unrelated to the scientific task (Adebayo et al., 2018). Methods of generalization may help to increase confidence in identified regions, but statistical guarantees often remain out of reach.

Furthermore, most recent problems associated with medical data have moved past simple difference detection: trends over time, and the ability to predict *future* disease development has by far become the setting of most interest. Given an image of a healthy individual, is it possible to predict what their scan, or their future disease diagnosis, may be up to 10, 20, or more years in the future? If a number of scans have been collected over some timeframe, can the *trajectory* of the individuals' development be extrapolated to estimate progression? As traditional models extended for temporal analysis grow in both size and complexity, a number of subproblems explicitly related to model and input subspaces arise. Here we aim to address two such problems: **statistically rigorous identification of temporally evolving subsets**, and **characterizations of deep models that enable efficient training of recurrent models with large scale time-varying data**.

With the rapid growth of AI and machine learning applications has come valid concerns regarding both guarantees of privacy. Recent technology legislation has made the importance clear in all aspects of data use, and particular projects and groups have demonstrated that machine learning is not independent of this need (Harvey, 2021). A new issue raised within this intersection is the "right to be forgotten". If a model has been trained with a particular users' data, they should have some recourse or right to both remove their data from the training set, and also know that the model has not learned from their data. On the surface, this poses a significant problem for model builders and organizations that spend large amounts of time and resources in training deep learning models. As we will see, **identification of model parameter subsets** that are particularly important for a particular sample's influence in a model enables *efficient machine unlearning*.

A sample's particular influence on model parameters aside, the identification of influential samples or subsets of samples more generally is of independent interest. Traditionally a rigorous area of study under classical statistics, outlier detection and accounting have become a subfocus for many within the machine learning community as well (Golatkar et al., 2020a,b; Huang et al., 2020; Ren et al., 2019). While subgroups of input samples may be outliers, it is more often the case that they represent known heterogeneity within the data. These differences are typically marked using group information known a priori, and most learning tasks aim to learn tasks in a *subgroup-independent* manner. Optimization and regularization methods with this focus come under the umbrella of model fairness, and instead of identifying and boosting in-

dependences within the model or data, we aim to minimize them. However, many existing methods do not scale well as the number of subgroups grows, as is often the case when intersections of protected classes must be considered. In the sequel we identify and construct a particular solution for **groupwise fairness that enables efficient in the loop fairness regularization**.

Thesis Goal: *Identify, construct, and evaluate methods for subset identification in machine learning input, feature, and model spaces, taking advantage of inherent geometric and statistical structures that naturally arise in application domains. Application focus is put on scientific domains in healthcare and social equity.*

1.1 Thesis Scope and Contributions

In this thesis we explore the intersections of classical statistical and geometric constructions with modern machine learning methods. Figure 1.1 shows the overall scope projected along three axes: feature, parameter, and sample spaces. Below we briefly introduce the main problems studied in this thesis.

Second-Order Modeling and Group Difference Analysis over Time

Recent results in coupled or temporal graphical models offer schemes for estimating the relationship structure between features when the data come from related (but distinct) longitudinal sources. A novel application of these ideas is for analyzing group-level differences, i.e., in identifying if *trends* of estimated objects (e.g., covariance or precision matrices) are different across disparate conditions (e.g., gender or disease). Often, poor effect sizes make detecting the *differential* signal over the *full* set of features difficult: for example, dependencies between only a *subset of features* may manifest differently across groups. We first suggest a parametric model for estimating trends in the space of SPD matrices as a function of one or more covariates. We will then generalize scan statistics to graph structures, to search over distinct subsets of features (graph partitions) whose temporal dependency structure may show statistically significant group-wise differences. We will theoretically analyze the Family Wise Error Rate (FWER) and bounds on Type 1 and Type 2 error. On a cohort of individuals with risk factors for Alzheimer's disease (but otherwise cogni-

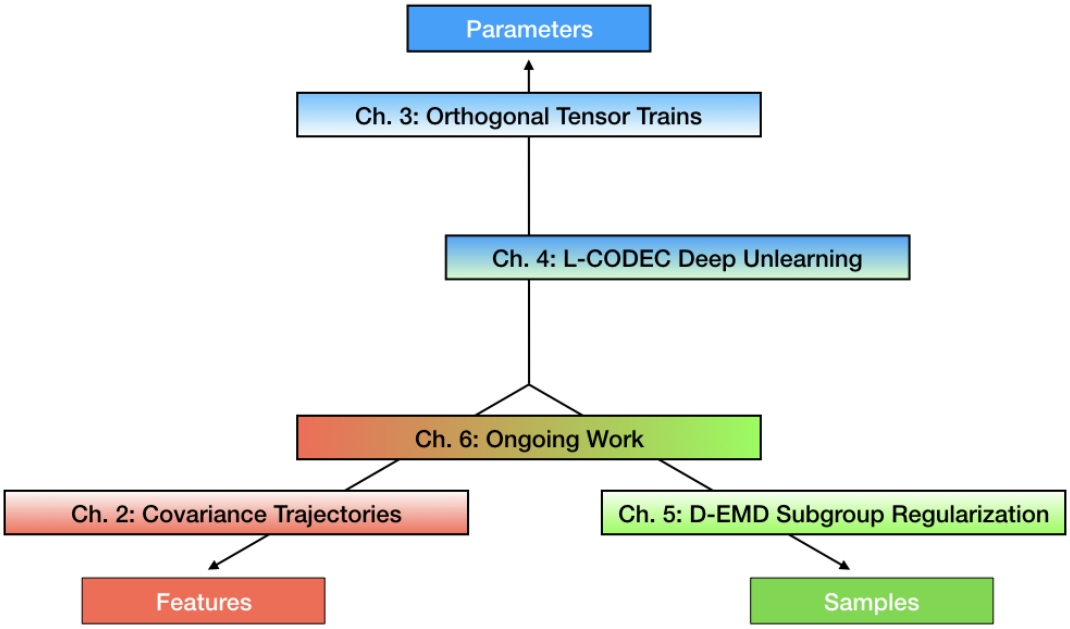

Figure 1.1: Thesis scope, projected over three representative axes.

tively healthy), we aim to find scientifically interesting group differences where the default analysis, i.e., models estimated on the *full graph*, do not survive reasonable significance thresholds. Preliminary work on this was published in (Mehta et al., 2019b).

Efficient Tensor Representations for Feasible Temporal Deep Learning

Modern deep networks have proven to be very effective for analyzing real world images. However, their application in medical imaging is still in its early stages, primarily due to the large dimension of three-dimensional images, requiring enormous convolutional or fully connected layers – if we treat an image (and not image patches) as a sample. These issues only compound when the focus moves towards longitudinal analysis through recurrent structures, and when a point estimate of model parameters is insufficient in scientific applications where a reliability measure

is necessary. Using insights from differential geometry, we will adapt the tensor train decomposition to construct networks with significantly fewer parameters, allowing us to train powerful recurrent networks on whole brain image volumes. We propose the *orthogonal tensor train*, and demonstrate its ability to express a standard network layer both theoretically and empirically. We will demonstrate its ability to effectively reconstruct whole brain volumes with faster convergence and stronger confidence intervals compared to the standard tensor train decomposition. We provide code and show experiments on the ADNI dataset using image sequences to regress on a cognition related outcome. Preliminary work on this was published in (Mehta et al., 2019a).

Practical Unlearning via Large-Scale Conditional Independence Testing

Recent legislation has led to interest in *machine unlearning*, i.e., removing specific training samples from a *predictive* model as if they never existed in the training dataset. Unlearning may also be required due to corrupted/adversarial data or simply a user's updated privacy requirement. For models which require no training (k-NN), simply deleting the closest original sample can be effective. But this idea is inapplicable to models which learn richer representations. Recent ideas leveraging optimization-based updates scale poorly with the model dimension d, due to inverting the Hessian of the loss function. We propose a variant of a new conditional independence coefficient, L-CODEC, to identify a subset of the model parameters with the most semantic overlap on an individual sample level. Our approach completely avoids the need to invert a (possibly) huge matrix. By utilizing a Markov blanket selection, we premise that L-CODEC is also suitable for deep unlearning, as well as other applications in vision. Compared to alternatives, L-CODEC makes approximate unlearning possible in settings that would otherwise be infeasible, including vision models used for face recognition, person re-identification and NLP models that may require unlearning samples identified for exclusion. Preliminary work on this will appear in (Mehta et al., 2022).

Reducing Subgroup Fairness via High Dimensional Earth Mover's Distances

Optimal transport has recently emerged as a useful tool for machine learning through its connections with geometry, statistical machine learning, and through practical algorithms. Existing methods that leverage optimal transport often regularize using a Wasserstein metric or by computing barycenters, for example. We will leverage optimal transport, except that we take advantage of a recently-introduced algorithm that computes a generalized earth mover's distance. Not only is this algorithm computationally cheaper to compute compared to existing barycentric measures, but our method has the additional advantage that gradients used for backpropagation can be directly read off of the forward pass computation, which leads to substantially faster model training. We will provide technical details about this new regularization term and its properties, and experimental demonstrations of improved training speed over existing Wasserstein-style methods.

Applying Conditional Independence Testing for better Understanding of Preclinical Alzheimer's Disease

The final chapter of this thesis applies some of the tools developed above in analyzing preclinical Alzheimer's disease patients. In these studies, we aim to identify conditionally independent features and subjects that are particularly important to the prediction and estimation of key disease outcomes, as a function of a number of demographic, neuropsychological, genetic, and imaging data collected as part of an ongoing consortium to understand the progression of Alzheimer's disease in younger, asymptomatic populations. In what follows we present exploratory analysis on a small, easily digestible subset of the available data, that lays the foundation for further analysis.

Outline

In Chapters 2 through 5, we describe four perspectives to address subset identification. Chapter 2 explores and focuses on the identification of feature subsets varying over time. In Chapter 3 we describe a method of constraining the parameter space in a particular manner that enables more efficient large scale neural networks. Next,

Chapter 4 provides a solution to the machine unlearning problem, enabled through a particular conditional independence parameter selection scheme, vastly reducing network update costs. Chapter 5 ends with a unique solution to subgroup fairness, where we take advantage of an efficient solution to the d-dimensional earth mover's problem to regularize large models when the number of subgroups can be large. Chapter 6 describes future work, focused on applying a particular solution from Chapter 4 to understanding relationships among disease indicators and biomarkers associated with developing Alzheimer's Disease.

Chapter 2

Localizing Group Differences over Covariance Trajectories

Feature selection has become a core problem with the increase in dimensionality of learning problems. Identifying and exploiting conditional independencies within data enables previously intractable problems to be solved, and this chapter focuses on one particular angle along this direction. When data are multivariate Gaussian, the zeros in the inverse covariance (precision) matrix give conditional independences among the variables (Lauritzen, 1996). Further, if the precision matrix is sparse, we can derive dependencies between features when the data are high-dimensional and/or the number of measurements are small. The estimation of a graphical model has been extensively studied and a rich literature is available describing its statistical and algorithmic properties (Koller and Friedman, 2009; Jordan, 1998). For instance, the so-called *graphical lasso* formulation uses an ℓ_1-norm penalty on the precision matrix and is widely used, and consistency properties in the large p regime (Cai et al., 2011; Friedman et al., 2008; Yuan, 2010) are now well understood. These formulations have also been extended to various transformations of Gaussian distributions (e.g., non-paranormal) using rank statistics (Liu et al., 2009; Xue and Zou, 2012; Liu et al., 2012).

Coupled and Temporal Graphical Models. Often, data come from two (or more) disparate sources or multiple timepoints. Within the last few years, a few proposals have described strategies for linking the sparsity patterns of multiple graphical models, e.g., using a fused lasso penalty (Danaher et al., 2014) (Yang et al., 2015). Observe that if the data sources correspond to *longitudinal* acquisitions, we should expect the

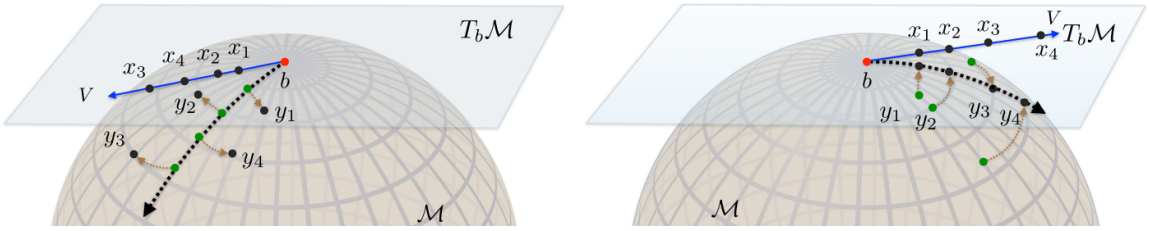

Figure 2.1: Group-wise MMGLM: The left and right figures represent two linear models on the SPD(p) manifold. Points x_i in the tangent space are our covariate or predictor, and points y_i in the manifold space represent SPD(p) matrices. In our regression setting, we wish to minimize the error (brown curves) between the estimation and the sample points. Because each linear model has a different base point, the trajectories cannot be directly compared as in the Euclidean setting.

'structure' to gradually evolve.

The ideas in the literature so far to "couple" multiple graphical model estimation modules are mostly nonparametric (Zhou et al., 2010; McArdle and Bell, 2000; Qiu et al., 2015). While such a formulation offers benefits, in many estimation problems, parametric models may be more convenient for downstream statistical analysis, particularly for hypothesis testing (Hardle and Mammen, 1993; Geer, 2000; Roehrig, 1988).

This will involve parameterizing *trends* in the highly structured nature of the 'response' variable (SPD matrices). Parametric formulations for manifold-valued data *have* been proposed recently (Kim et al., 2014; Cornea et al., 2016). Because SPD matrices form a Riemannian manifold, algorithms that estimate a parametric model respecting the underlying Riemannian metric are more suitable in many applications as opposed to assuming a Euclidean metric on positively or negatively curved spaces (Xie et al., 2010; Fletcher and Joshi, 2007; Jayasumana et al., 2013). We will make a few simple modifications (for efficiency purposes) to such algorithms and make use of the estimated parameters for follow-up analysis.

Finding Group-wise Differences. Assuming that we have a black-box procedure to estimate a parametric model on the SPD manifold available, in many tasks, such an estimation is merely a segue to other analyses designed to answer scientifically meaningful questions. For example, we are often interested in asking whether the temporally coupled model estimated using the procedure above differs in meaningful ways *across* groups induced by a stratification or dichotomous variable (e.g., gender or disease). For instance, is the 'slope' in structured response space statistically different

across education level or body mass index? While the body of work for graphical model estimation is mature, the literature describing hypothesis tests in this regime (Städler and Mukherjee, 2012; Belilovsky et al., 2015) is sparse at best.

Given that such questions are simpler to answer with alternative schemes (with assumptions on the distributional properties of the data), e.g., structural equation modeling, latent growth models and so on (Ullman and Bentler, 2003; McArdle and Bell, 2000), it seems that the unavailability of such tools is limiting the adoption of such ideas in a broader cross-section of science. We will seek to address this gap.

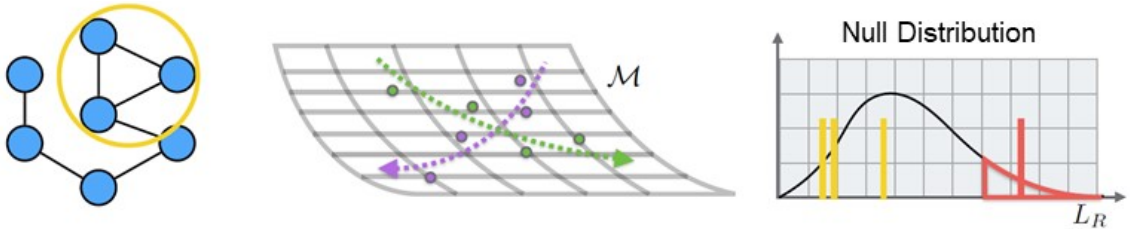

Figure 2.2: Our proposed method involves (1) electing a subset of features, (2) fitting manifold regressions on empirical covariances for both groups (green and purple) and (3) constructing the likelihood ratio statistic and comparing it against the null distribution via permutation testing.

Needles in Temporal Haystacks. If we temporarily set aside the potential value of a hypothesis test framework for temporal trajectories in graphical models, we see that from an operational viewpoint, such procedures are most effective when a practitioner already has a precise scientific question in mind. In reality, however, many data analysis tools are deployed for exploratory analyses to inform an investigator as to which questions to ask. Being able to "localize" which parts of the model are different across groups can be very valuable. This ability actually benefits statistical power as well. Notice that when the stratified groups are not very different to begin with, e.g., healthy individuals with presence or absence of a genetic mutation, the effect sizes (statistical difference between two groups) are likely to be poor. Here, while the trends identified on the *full* precision matrix may still be different (i.e., there may be a *real* signal associated with a grouping variable), they may not be strong enough to survive significance thresholds. Ideally, what we need here are analogs of the widely used "scan statistics" for our hypothesis testing formulations for temporal graphical models — to identify which *parts of the signal* are promising. Then, even if only a small subset of features were different across groups, we may be able to identify these differential effects efficiently. This benefits Type 2 error, provides a practical

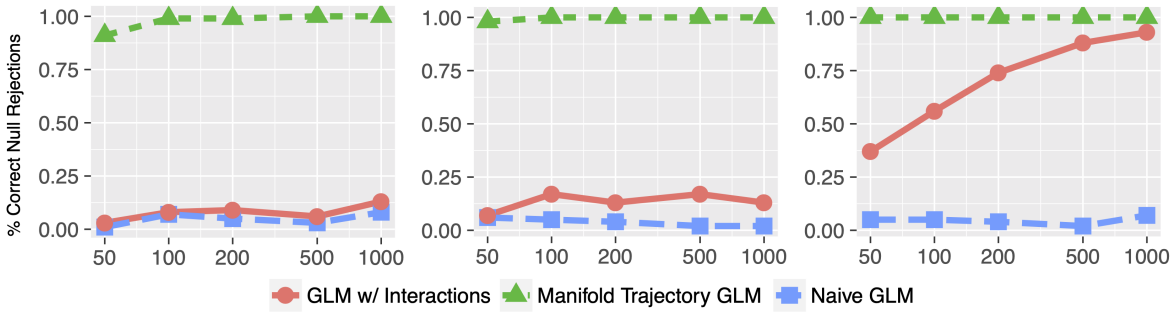

Figure 2.3: Correct null hypothesis rejections over 100 runs for three models. For $p = 50$ features, each plot shows the rejection rate for $p_t \in \{4, 8, 20\}$ (from left to right) respectively as a function of the number of sample points.

turnkey product for an experimental scientist, and makes up the key technical results of our work.

Simulations. We randomly generate SPD matrices from a 'path' of 4 discrete points along the manifold, and use these data as population covariance matrices to generate 0-mean sample data. We compare our model to baseline methods that may be used in practice for a group difference hypothesis test. In standard applications, general linear models (GLMs) are often the first line of attack. When the covariates are assumed to be independent, a simple linear model may be suitable. However, when the group difference is influenced by specific interactions between covariates, such linear models require additional care. A typical solution is to introduce pairwise interaction terms into the model – a choice between all possible interactions or *specific interactions specified by an expert*. The first model has problems since the number of samples $n \ll p^2$. In the second model, we depend completely on the user's choice of interactions, and must correct for multiple testing when testing different models, at least partly reducing the power of the final test. Figure 2.3 shows the value of our method over these models. For the interaction GLM case, we randomly select interaction terms to include in the GLM, with size p_t (the ground truth number of variables in the interaction). In this way, we approximate the effect of an oracle specifying to the GLM which terms may describe the underlying interaction. We report the fraction of significance tests where a significance threshold of $p \leqslant 0.05$ was found for each model, averaged over 100 runs. We see that our proposed scheme consistently achieves near-perfect results in terms of the percentage of null hypotheses that were correctly rejected (i.e., there was a significant group-difference signal). The

power of scan statistics on graphs is particularly evident in the needle in haystack setting where the true differential signal is small ($p_t \leqslant 8$) and the sample size is small to medium. When the sample size is large and p_t is also large, the standard linear model with additional interaction terms starts to approach the statistical performance of our algorithm.

Briefly, we will provide **(i)** a simple and efficient parametric procedure for modeling temporally evolving graphical models, **(ii)** a hypothesis test for identifying differences between group-wise estimated models, and **(iii)** a scan algorithm to identify *those subsets of the features which contribute to the group-wise differences*. Together, these ideas offer a framework for identifying group-wise differences in temporally coupled graphical models. From the experimental perspective, we hope to find scientifically plausible results on a unique longitudinally tracked cohort of middle-aged (and young elderly) persons at risk for Alzheimer's disease due to family history, but who are otherwise completely cognitively healthy.

Chapter 3

Enabling Temporal Neural Networks via Geometric Tensor Representations

While feature selection dates back to classical statistics, *parameter selection* has only recently been studied as the dimensionality and complexity of models grows linearly with the size of modern neural networks. Recurrent Neural networks (RNNs) and its variants are the de facto tool of choice for modeling sequential data in machine learning and vision. But until only recently, these models have been severely limited in their ability to model high-dimensional data. Recurrent structures often lead to large model sizes dependent on sequence length, and thus also require an equivalent number of increased computation. While RNNs have been successfully applied to video data in some cases, the strategy requires problem specific innovations because of the large mapping necessary from inputs to hidden representations. It is fair to say that the growth in the number of model parameters in various types of recurrent models remains a serious bottleneck for high dimensional datasets. For model-size reduction, both for RNN style networks and otherwise, PCA or random projections (Ye et al., 2005; Bingham and Mannila, 2001) style “compression" ideas have also been used with varying degrees of success.

An interesting perspective on the effective degrees of freedom afforded by a given network, a surrogate for the actual “size” of the architecture, is provided by tensor methods. Tensor decomposition based methods have recently been shown to enable low dimensional representations of very high dimensional data, and while these ideas were known to be effective in the “shallow" regime much earlier, new results also demonstrate their applicability for deep neural networks. In particular, in the last year,

we see a number of tensor based methods being successfully adapted for deep neural network design and compression (Cohen et al., 2016; Zhang et al., 2017; Yu et al., 2017). Specifically, (Yang et al., 2017) shows that these compression methods can be very effective in reducing the parameter cost of weight layers in RNNs, enabling simple video analysis tasks that previously would have been computationally prohibitive.

Our goal is to design rich sequential or recurrent models to analyze a longitudinal sequence of high dimensional 3D brain images. This task raises two major issues. **First**, unless the model size is parsimonious, we find that merely instantiating the model with data involving 3D images over multiple time points, even on multiple high end GPU instances, is challenging. **Second**, the eventual goal of medical image analysis is either scientific discovery or generating actionable knowledge for patient betterment. Both goals require evaluating a model's confidence via classical or contemporary statistical techniques: for instance, how confident is the model of its prediction? Most, if not all, available tools for assessing model uncertainty of deep neural network models have a strong dependence on the number of parameters in the model. Therefore, even if the first issue above could be mitigated by clever implementation ideas, purely as a practical matter, the design of rich and expressive models with a small number of parameters yields immense benefits for calculating model uncertainty.

The Central Idea. Let $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be a d-dimensional array, or tensor, with each mode having length n_i. To store a full rank tensor, n^d storage would be required. A number of tensor factorizations have been developed to reduce this storage cost. The CANDECOMP/PARAFAC (CP) (Harshman, 1970; Carroll and Chang, 1970) decomposition reduces the storage to $O(dnr)$, but finding the exact CP-rank r is NP-hard. Hierarchical tensor methods have also proven to be effective in tensor compression (Cohen et al., 2016; Cohen and Shashua, 2016).

The *Tensor Train* decomposition (TT) (Oseledets, 2011), defines an element of the tensor as

$$\mathcal{X}(x_1, \ldots, x_d) = A_1(x_1) \cdots A_d(x_d) \tag{3.1}$$

where $x_i \in \{1, \ldots, n_i\}$, and $A_i(x_i) \in \mathbb{R}^{r_{i-1} \times r_i}$ for each $i \in \{1, \ldots, d\}$ are called the *cores*

of the tensor train, with $r_0 = r_d = 1$. Equivalently, the full tensor is written as:

$$\mathcal{X} = \sum_{k_0=1}^{r_0} \cdots \sum_{k_d=1}^{r_d} A_1(k_0, :, k_1) \otimes \cdots \otimes A_d(k_{d-1}, :, k_d) \tag{3.2}$$

where $A_i \in \mathbb{R}^{r_{i-1} \times n_i \times r_i}$. This format requires $O(dnr^2)$ storage, but has two major advantages over the CP format. First, finding the TT-rank (the smallest set of r_i's that satisfy the decomposition with equality) of any arbitrary tensor is tractable, and as such all tensors can be efficiently rewritten in the TT format. Second, projecting arbitrary tensors onto the TT format of a fixed rank requires only a set of QR and singular value decompositions (Oseledets, 2011). This projection, *TT-rounding*, additionally allows for a given TT tensor of some rank to be projected onto the space of TTs with lower rank, and requires $O(dr^3)$ computational complexity. Separately, specific tensor train constructions have recently been identified as forms of general recurrent networks (Khrulkov et al., 2019)

Ideally, we would prefer a construction which keeps the standard TT-core format and involves optimization over "smaller" Stiefel manifolds. Consider the following representation, in which each TT-core itself is orthogonal.

Definition 3.1. (*Orthogonal Tensor Train*) *The Orthogonal Tensor Train is defined as*

$$\mathcal{X}(x_1, \ldots, x_d) = Q_1(x_1) \cdots Q_d(x_d), \tag{3.3}$$

where each $Q_i(x_i)$ *lies on the Stiefel* $St(m_i, M_i)$, *where* $m_i = \min(r_{i-1}, r_i)$, *and* $M_i = \max(r_{i-1}, r_i)$.

While in this formulation the total number of components in the product space of Stiefels is nd, the dimension of each manifold is **significantly smaller**, dependent *only* on the core rank as opposed to the mode size. The total number of parameters, if $n_i = n, r_i = r$, is

$$n \sum_{i=1}^{d} \left[r^2 - \frac{r^2 + r}{2} \right] = dnr^2 - dn\frac{r^2 + r}{2}. \tag{3.4}$$

When compared to the full TT representation, the Orthogonal Tensor Decomposition (OTT) requires $(r+1)/2r \approx 1/2$ as many parameters. If $r_i = r_{i+1}$, then $St(m_i, M_i) = SO(m_i)$, where SO is the special orthogonal group.

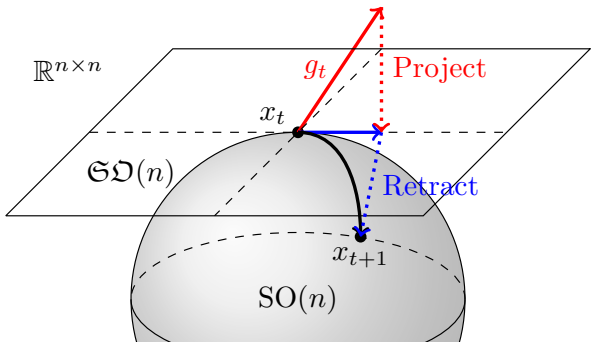

Figure 3.1: Visualization of the gradient descent update using the projection and retraction on the Stiefel manifold. The update is applied to each core individually, allowing for smaller manifold operations that would otherwise scale poorly with dimension.

This construction can be seen as an approximation to the full tensor train format, in which the upper triangular part of each core is set to identity:

$$\begin{aligned}\mathcal{X}(x_1,\ldots,x_d) &= A_1(x_1)\cdots A_d(x_d)\\ &= Q_1(x_1)R_1(x_1)\cdots Q_d(x_d)R_d(X_d)\\ &\approx Q_1(x_1)\cdots Q_d(x_d)\end{aligned}$$

Stochastic OTT Optimization

for t=1,...,T **do**
  $g_t := \frac{df}{dW} f(X^{mini-batch})$
  for Core $Q_t^i \in \mathcal{W}_t$ and Core Gradient $g_t^i \in g_t$ **do**
    $G_t^i = P_{T_{\mathcal{W}_t}\mathcal{M}}(g_t^i)$ ▷ Projection Step
    $Q_{t+1}^i \leftarrow \mathrm{Exp}(Q_t^i, G_t^i)$ ▷ Retraction Step
  end for
end for

Goals. Using this construction, we will tackle the problem of modeling sequential 3D brain imaging data using recurrent/sequential models. Our development starts from well known results on tensor decomposition, and in particular, we make use of the tensor train representation, which has been shown to be effective in several applications in vision and machine learning. We derive a reformulation of the decomposition using orthogonality constraints and show that while this makes the estimation slightly more challenging, it reduces the number of parameters by as much as half. We present a novel parameter estimation scheme based on Stiefel manifold optimization and demonstrate how the end to end construction yields benefits for

convergence and uncertainty estimation. Finally, from the empirical side, we will discuss how we enable analysis of and prediction using sequential 3D brain imaging datasets, which to our knowledge is the first such result using deep recurrent/sequential architectures.

Chapter 4

Efficient Learning and Unlearning via Large-Scale Conditional Independence Testing

While from one perspective we may be able to constrain our parameter space to one which is desirable, it may be the case that we cannot affect or intervene in the model prior to training. In these cases, we may still wish to identify a *subset of existing parameters* that are important or related to particular samples or sample subsets. As personal data becomes a valuable commodity, legislative efforts have begun to push back on its widespread collection/use particularly for training ML models. Recently, a focus is the "right to be forgotten" (RTBF), i.e., the right of an individual's data to be deleted from a database (and derived products). Despite existing legal frameworks on fair use, industry scraping has led to personal images being used without consent, e.g. (Harvey, 2021). Large datasets are not only stored for descriptive statistics, but used in training large models. While regulation (GDPR, CCPA) has not specified the extent to which data must be forgotten, it poses a clear question: is deletion of the data enough, or does a model trained on that data also needs to be updated?

Recent work by (Carlini et al., 2019, 2020) has identified scenarios where trained models are vulnerable to attacks that can reconstruct input training data. More directly, recent rulings by the Federal Trade Commission FTC (2021); Kaye (2022) have ordered companies to fully delete and destroy not only data, but also any model trained using those data. While deletion and (subsequent) full model retraining without the deleted samples is possible, most in-production models require weeks

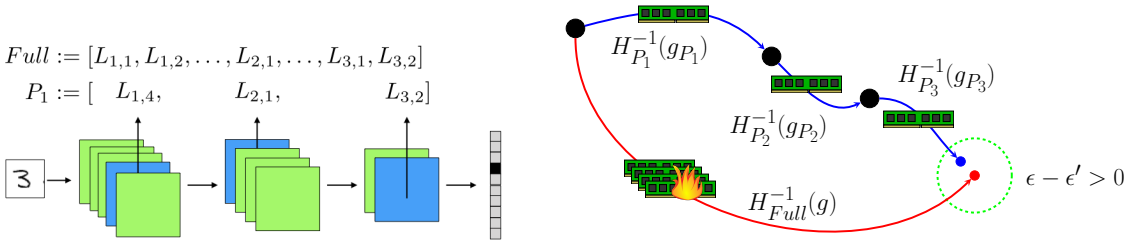

Figure 4.1: Large deep learning networks typically associate specific subsets of network parameters, blocks (blue), to specific samples in the input space. Traditional forward or backward passes may not reveal these blocks: high correlations among features may not distinguish important ones. Input perturbations can be used to identify them in a probabilistic, distribution-free manner. These blocks can then be unlearned together in an efficient block-coordinate style update (right, blue lines), approximating an update to the full network which requires a costly/infeasible full Hessian inverse (red line).

of training and review, with extensive computational/human resource cost. With additional deletions, it is infeasible to retrain each time a new delete request comes in. So, how to update a model ensuring the data is deleted without retraining?

Task. Given a set of input data $\mathcal{S} : \{z_i\}_{i=1}^n \sim \mathcal{D}$ of size n, training simply identifies a hypothesis $\hat{w} \in \mathcal{W}$ via an iterative scheme $w_{t+1} = w_t - g(\hat{w}, z')$ until convergence, where $g(\cdot, z')$ is a stochastic gradient of a fixed loss function. Once a model at convergence is found, *machine unlearning* aims to identify an update to $\hat{w}$ through an analogous *one-shot unlearning update*:

$$w' = \hat{w} + g_{\hat{w}}(z'), \tag{4.1}$$

for a *given* sample $z' \in \mathcal{S}$ that is to be **unlearned**.

Let $\mathcal{A}$ be an algorithm that takes as input a training set $\mathcal{S}$ and outputs a hypothesis $w \in \mathcal{W}$, defined by a set of d parameters Θ. An unlearning scheme $\mathcal{U}$ takes as input a sample $z' \in \mathcal{S}$ used as input to $\mathcal{A}$, and ideally, outputs an **updated** hypothesis $w' \in \mathcal{W}$ where z' has been deleted from the model. An unlearning algorithm should output a hypothesis that is close or equivalent to one that would have been learned had the input to $\mathcal{A}$ been $\mathcal{S} \setminus z'$. A framework for this goal was given by Ginart et al. (2019) as,

Definition 4.1 ($(\epsilon, \delta)-$ forgetting)**.** *For all sets $\mathcal{S}$ of size n, with a "delete request" $z' \in \mathcal{S}$,*

an unlearning algorithm $\mathcal{U}$ is (ϵ, δ)—forgetting if

$$\mathbb{P}(\mathcal{U}(\mathcal{A}(S), z') \in \mathcal{W}) \leqslant e^{\epsilon}\mathbb{P}(\mathcal{A}(S \setminus z') \in \mathcal{W}) + \delta \tag{4.2}$$

In essence, for an existing model w, a good unlearning algorithm for request $z' \in S$ will output a model $\hat{w}$ close to the output of $\mathcal{A}(S \setminus z')$ with high probability.

Remark 4.2. *Definition 4.1 is similar to the standard definitions of differential privacy. The connection to unlearning is: if an algorithm is (ϵ, δ)—forgetting for unlearning, then it is also differentially private.*

If $\mathcal{A}$ is an empirical risk minimizer for the loss f, let

$$\mathcal{A} : (S, f) \to \hat{w} \tag{4.3}$$

$\hat{w} = \arg\min F(w)$ and $F(w) = \frac{1}{n}\sum_{i=1}^{n} f(w, z_i)$. Recall $g(z')$ from (4.1): our unlearning task essentially involves identifying the form of $g(z')$ for which the update in (4.1) is (ϵ, δ)-forgetting. If an oracle provides this information, we have accomplished the unlearning task.

The difficulty, as expected, tends to depend on f and $\mathcal{A}$. Recent unlearning results have identified forms of f and $\mathcal{A}$ where such a $g(z')$ exists. The authors in (Sekhari et al., 2021) define $g(z') = \frac{1}{n-1}H'^{-1}\nabla f(\hat{w}, z')$, where

$$H' = \frac{1}{n-1}\left(n\nabla^2 F(\hat{w}) - \nabla^2 f(\hat{w}, z')\right), \tag{4.4}$$

with additive Gaussian noise $w' = w' + N(0, \sigma^2)$ scaling as a function of n, ϵ, δ, and the Lipschitz and (strong) convexity parameters of the loss f. We can interpret the update using (4.4) from the optimization perspective as a trajectory "reversal": starting at a random initialization, the first order (stochastic gradient) trajectory of w (possibly) *with* z' is reversed using *residual* second order curvature information (Hessian) at the optimal $\hat{w}$ in (4.4), achieving unlearning. This is shown to satisfy Def. 4.1, and only incurs an additive error that scales by $O(\sqrt{d}/n^2)$ in the gap between $F(w')$ and the global minimizer $F(w^*)$ over the ERM $F(\hat{w})$.

Rationale for approximate schemes. From the reversal of w optimization perspective, it is clear that there may be other choices to achieve unlearning. For a practitioner interested in unlearning, the aforementioned algorithm (as in (4.4)) can be directly

instantiated if one has extensive computational resources. Indeed, in settings where it is not directly possible to compute the Hessian inverse necessary for $H'^{-1}\nabla f(\hat{w}, z')$, we must consider alternatives.

A potential idea. Our goal is to identify a form of $g(z')$ that **approximates** $H'^{-1}\nabla f(\hat{w}, z')$. Let us consider the Newton-style update suggested by (4.4) as a smoothing of a traditional first order gradient step. The inverse Hessian is a weighting matrix, appropriately scaling the gradients based on the second order difference between the training set mean point $F(\hat{w})$ and at the sample of interest $f(\hat{w}, z')$. This smoothing can also be seen from an information perspective: the Hessian in this case corresponds to a Fisher-style information matrix, and its inverse as a conditional covariance matrix (Golatkar et al., 2021, 2020b). It is not hard to imagine that from this perspective, if there are *specific set of parameters* that have *small gradients* at $f(\hat{w}, z')$ or if the information matrix is zero or small, then we need not consider their effect.

Contributions. We will address several computational issues with existing approximate formulations for unlearning by taking advantage of a new statistical scheme for sufficient parameter selection. First, in order to ensure that a sample's impact on the model predictions is minimized, we propose a measure for computing conditional independence called L-CODEC which identifies the Markov Blanket of parameters to be updated. Second, we will show that the L-CODEC identified Markov Blanket enables unlearning in previously infeasible deep models, scaling to networks with hundreds of millions of parameters. Finally, we will demonstrate the ability of L-CODEC to unlearn samples and entire classes on networks, from CNNs/ResNets to transformers, including face recognition and person re-identification models.

Chapter 5

Generalizing the Earth Mover's Distance for Efficient Neural Network Regularization

Selections of parameters related to individual samples can help with single sample problems, but the majority of learning problems are more concerned with generalization: if a subset of samples known to be related are behaving differently, can we identify them, and can we remove this difference? Here, we aim to account for groups of samples that behave differently compared to others, and efficiently regularize models toward fairer solutions.

Optimal transport (OT) has emerged as useful tool for a wide range of applications including information retrieval (Balikas et al., 2018; Yurochkin et al., 2019), image processing (Bonneel et al., 2014), statistical machine learning, as well as more recently, for applications in ethics and fairness (Kwegyir-Aggrey et al., 2021). OT is especially well-suited for tasks where dissimilarity between two or more probability distributions must be quantified, and its success has been made possible through dramatic improvements in numerical algorithms (Cuturi, 2013; Solomon et al., 2015) that allow one to efficiently optimize commonly used functionals. In practice, OT is often used to estimate and minimize the the distance between distributions of interest, and this is done using an appropriately defined loss functional.

Barycenters. One way to quantify dissimilarity between many distributions is through distance to the mean. Here, for averaging, we measure pairwise distances. Practically, this has led to models that can concurrently enforce distributional similar-

ity between a set of distributions and has found use in a spectrum of applications.

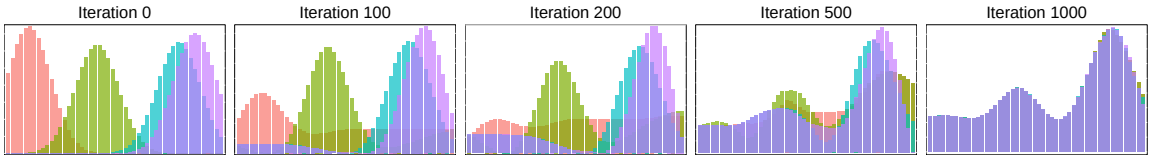

Figure 5.1: This figure shows the starting and ending state of an iterative process applied to 4 histograms. Each iteration minimizes the generalized Earth Mover's objective, and then updates each histogram in the direction provided by the gradient. Upon each iteration shifts a portion of "mass" of every histogram closer to a common distribution. (*Left*) The process initially starts with $d = 4$ histograms with 50 bins each. (*Right*) The update process eventually converges. An intuitive description is that the mass of the original 4 histograms is incrementally shifted around at each step, and all four distributions eventually are reshaped into the limiting distribution, which appears at right.

Assuming that a suitably regularized form of the optimal transport loss is utilized, the pairwise distance calculation is efficient – in fact, in some cases, Sinkhorn iterations can be used (Cuturi, 2013). On the other hand, to minimize distances to the mean, most algorithms typically operate by repeatedly estimating the barycenter and those pairwise distances, and using a "coupling" strategy to push points toward the barycenter. The idea is sensible but as the number of distributions grows, the overall procedure becomes computationally intensive. For example, even on a high end workstation, simply computing the barycenter over 50 distributions with 50 bins proves to be a time-intensive process. However, very recent research has shown that there exist polynomial time algorithms for this problem (Altschuler and Boix-Adsera, 2021).

Where can we measure dissimilarity? In applications, it is often a goal to enforce distribution similarity on model outputs. For example, in (Jiang et al., 2020), the authors define fairness measures over the probability of the prediction given ground truth labels. However, these methods are rarely extended to continuous measures within the neural network, mainly due to the strong distributional assumptions needed and the added algorithmic complexity of estimating the barycenter. One drawback of this choice means that the distributional closeness is only guaranteed over the global model output, and little can be said about layer outputs prior to the thresholding for prediction. Additionally, full retraining would be necessary if thresholds must be adjusted due to changing business or regulatory requirements.

We will use the following measure as our "dissimilarity." The generalized Earth Mover's Distance is

$$\begin{aligned}
&\underset{x\in\mathbb{R}^{n^d}}{\text{minimize}} \sum_{i_1,\ldots,i_d} c(i_1,\ldots,i_d)x(i_1,\ldots,i_d)\\
&\text{subject to}\\
&\sum_{i_2,\ldots,i_d} c(i_1,\ldots,i_d)x(i_1,\ldots,i_d) = p_1(i_i),\ (\forall i_1 \in [n])\\
&\sum_{i_1,i_3\ldots,i_d} c(i_1,\ldots,i_d)x(i_1,\ldots,i_d) = p_2(i_2),\ (\forall i_2 \in [n])\\
&\quad\vdots\\
&\sum_{i_1,\ldots,i_{d-1}} c(i_1,\ldots,i_d)x(i_1,\ldots,i_d) = p_d(i_d),\ (\forall i_d \in [n])
\end{aligned} \tag{5.1}$$

The solution to this program effectively identifies a joint distribution x such that the marginals satisfy all p_j, and which minimizes the cost. Although the optimal value of the objective function of this program no longer defines a metric when $d > 2$, it still possesses several desirable properties, which we use.

We can also write the dual linear program.

$$\begin{aligned}
&\underset{z_j\in\mathbb{R}^n, j\in[d]}{\text{maximize}} \quad \sum_j x_j' z_j\\
&\text{subject to} \quad z_1(i_1) + \cdots + z_d(i_d) \leqslant c(i_1,\ldots,i_d),
\end{aligned} \tag{5.2}$$

where the indices in the constraints include all $i_j \in [n]$, $j \in [d]$.

Our core observation revolves around the identification of the gradient that falls naturally from the construction in (Kline, 2019).

Theorem 5.1. *The following two claims hold. First,*

$$\nabla\phi(p_1,\ldots,p_d) = z^*$$

and second, for any $t \in \mathbb{R}$,

$$\phi(p_1, p_2, \ldots, p_d) = \sum_j p_j'(z_j^* + t\,\eta),$$

where $\eta \coloneqq (z_1^*(n)\,e, z_1^*(n)\,e, \cdots, z_d^*(n)\,e)$.

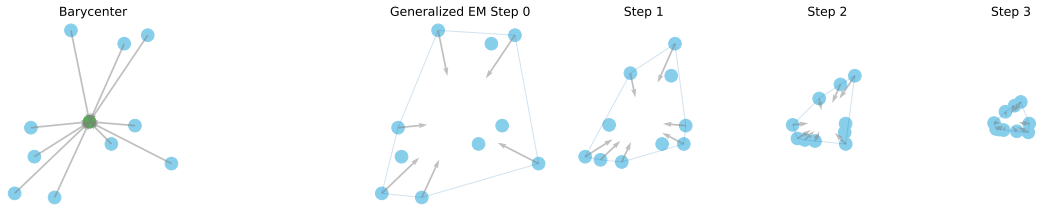

Figure 5.2: (*Left*) Barycenter approaches identify a center (green) and move samples of interest (blue) toward it along the coupling path (grey). (*Right*) Our approach identifies "support" points that lie within the convex hull, and only those points are moved in a descent direction. The support points are those that contribute to the objective functional, and it is known (see Theorem 2.2 of (Kline, 2019)) that they are not interior points of the convex hull.

Contributions. We exploit a recent extension of the classical Earth Movers Distance (EMD) to a higher-dimensional Earth Mover's objective. We will show that minimization of this *global* distributional measure leads to the harmonization of input distributions similar in spirit to the minimization of distributions to barycenters. We prove theoretical properties of the objective and our procedure, and show that the gradient can be read directly off from a primal/dual algorithm, alleviating the need for computationally intense pairwise couplings. With a particular scaffolding provided by differentiable histograms, we can apply and smoothly operate directly on network activations to compute the EMD measure. We will establish through experiment that computing gradients used in backpropogation is possible in substantially shorter times than one can achieve using standard tools, due to rapid access to solutions of the the dual linear program formulation. We will compare and contrast the performance and speed of our construction against Barycenter-like measures in a number of settings and demonstrate applications in a common fairness application. Our final construction integrates seamlessly with existing neural network pipelines, which will be publicly available for use.

Chapter 6

Ongoing Work: Large Scale Analysis of Multi-Site Preclinical Alzheimer's Disease

With the above tools in hand, we aim to move towards applying and deploying both conditional independence schemes and efficient EMD-fairness methods to the analysis of a new preclinical cohort of individuals at risk for developing Alzheimer's disease.

Data. Here our data consists of patient information pooled across multiple sites. Demographic measures, neuropsychological test results, genetic indicators, and cerebrospinal fluid (CSF) biomarkers were all collected on individuals from three studies: the Adult Children Study (ACS), the Wisconsin Registry for Alzheimer's Prevention (WRAP), and the Biomarkers of Cognitive Decline Among Normal Individuals (BIOCARD). Data was preprocessed using standard pipelines, collated, and harmonized across sites. Table 6.1 details the full list of measures.

Aβ42	Aβ40
T-Tau	P-Tau
Aβ42/Aβ40	PTAU-Aβ42
Age	Gender
Executive Function	General Cognitive Performance
Episodic Memory	Time
Education	APOE

Table 6.1: Preclinical AD Measures used in conditional independence analysis.

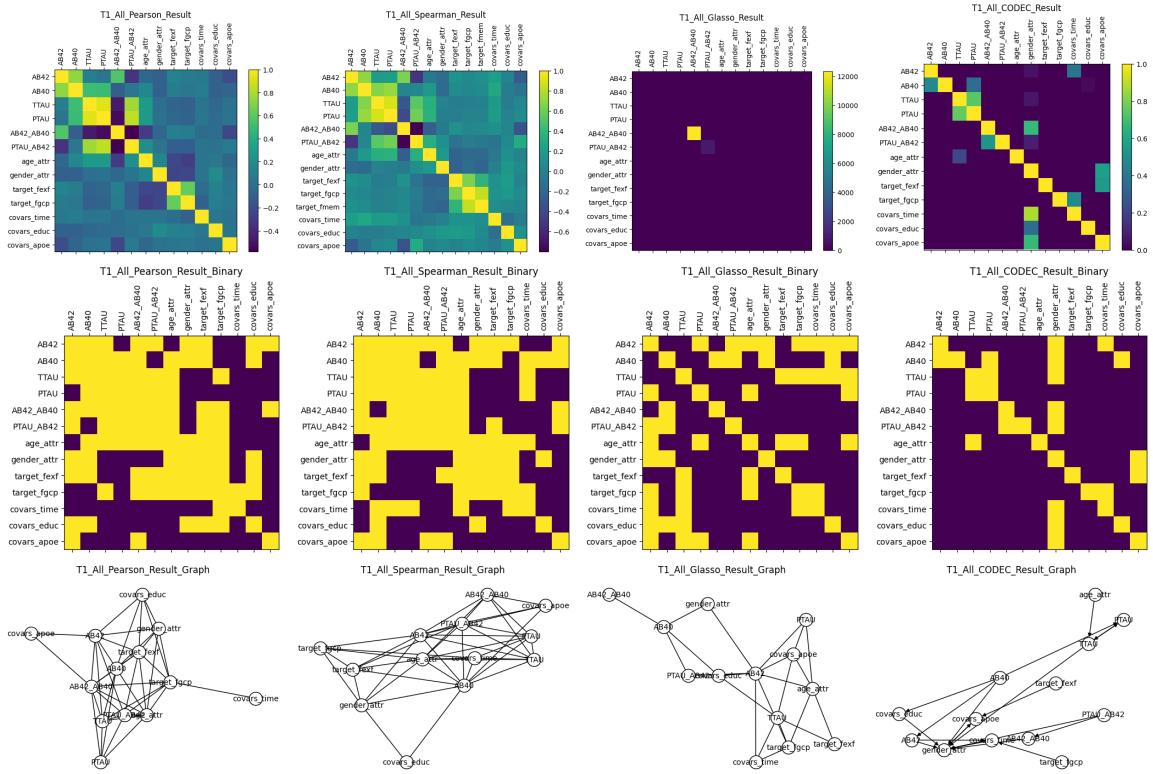

Figure 6.1: All Sites.

Imaging data is also available for a subset of the individuals included in the above set, and future work includes incorporation of these imaging modalities (MRI, DTI, and PET region of interest measures).

Preliminary Methods and Results. The results shown here demonstrate the potential value of applying conditional independence methods over standard correlation methods. The value of sparsity patterns derived from conditional independence testing appears to be clear, in that hyper-parameters need not be chosen in any way, compared to an α-level for hypothesis testing via correlation coefficients, or a regularization level λ for a typical graphical LASSO setting. Running the same analysis separately for each site, first observations indicate that there may exist dependencies and independencies unique to each study.

Upcoming Analysis. Future analysis will be focused on exploring the differences observed in the sparsity and conditional independences across different studies within the consortium data, as well as applying the construction longitudinally. Particular care will be taken in defining discrete timepoints, as disease progression has no

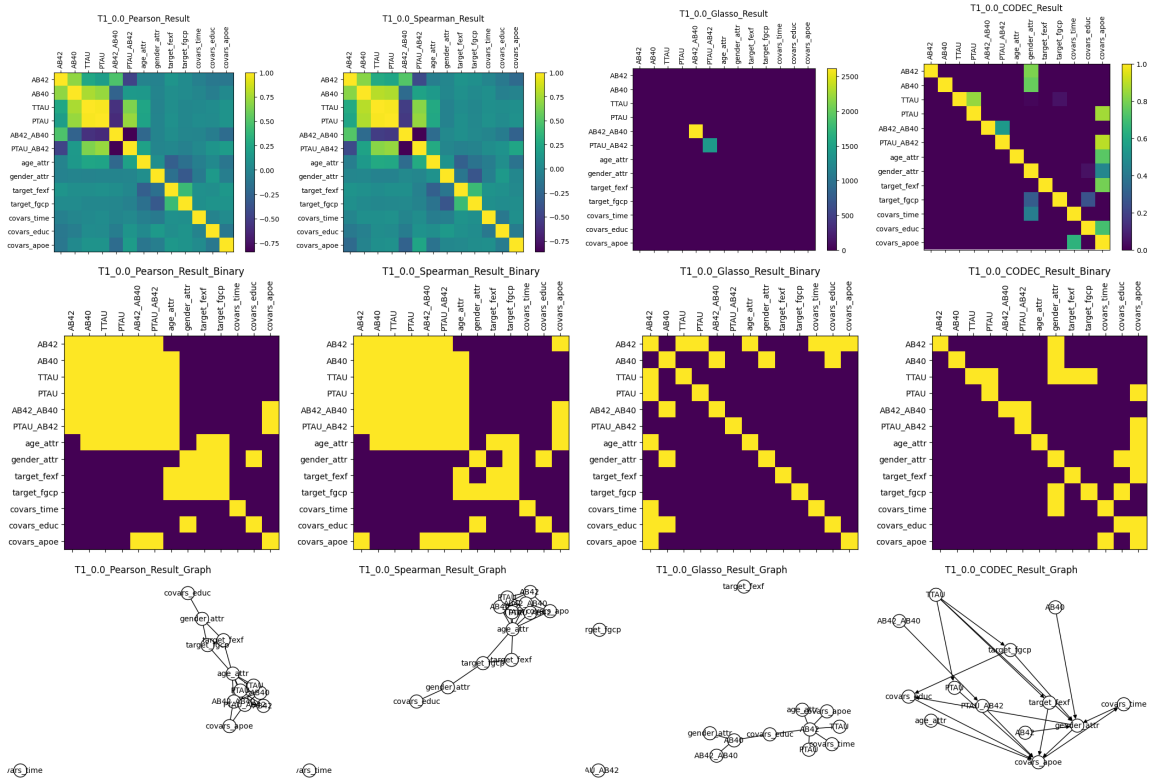

Figure 6.2: Site 0: WRAP.

predefined domain measure, and individuals within the study may be different ages at study times, as well as visiting at irregular intervals.

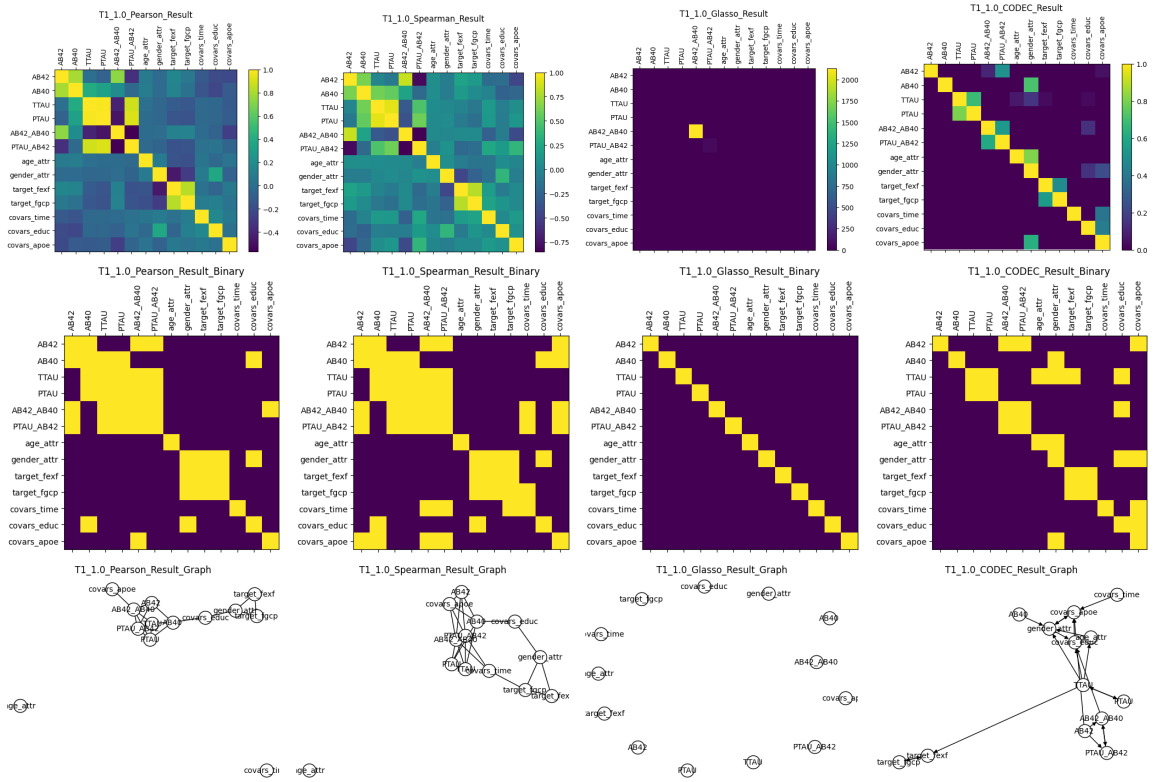

Figure 6.3: Site 1: ACS.

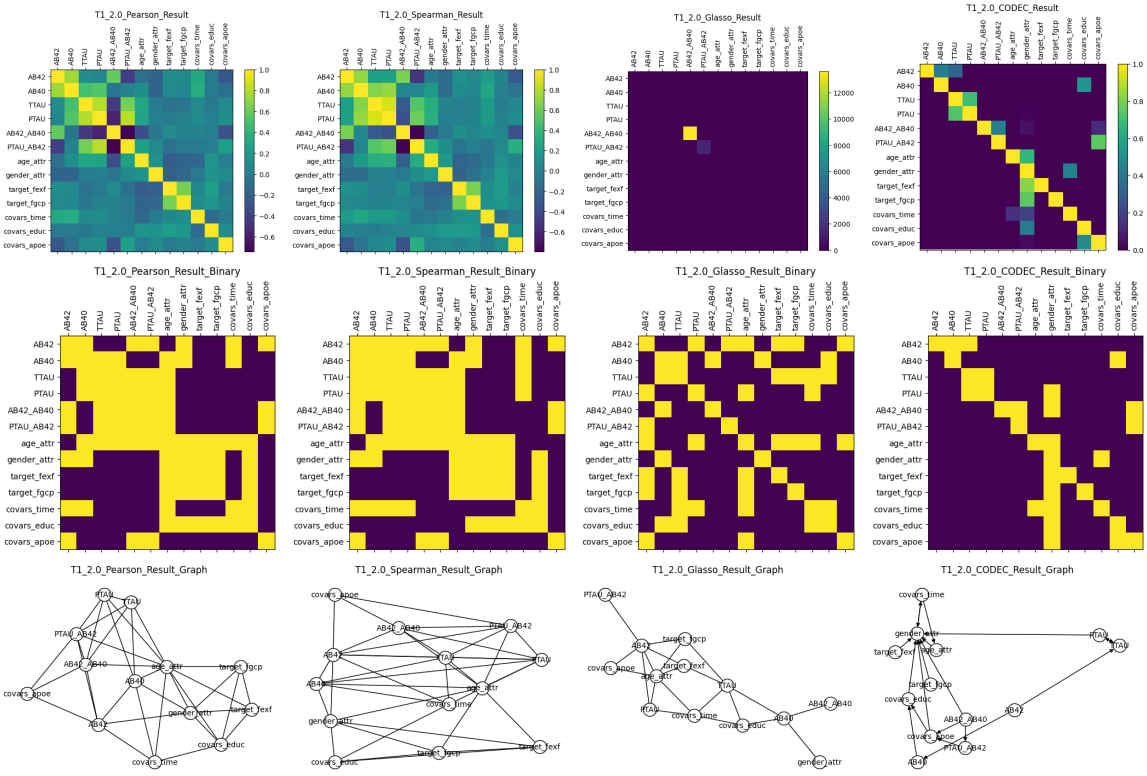

Figure 6.4: Site 2: BIOCARD.

references

Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems* 31.

Altschuler, Jason M, and Enric Boix-Adsera. 2021. Wasserstein barycenters can be computed in polynomial time in fixed dimension. *J. Mach. Learn. Res.* 22:44–1.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Balikas, Georgios, Charlotte Laclau, Ievgen Redko, and Massih-Reza Amini. 2018. Cross-lingual document retrieval using regularized wasserstein distance. In *European conference on information retrieval*, 398–410. Springer.

Belilovsky, Eugene, Gaël Varoquaux, and Matthew B Blaschko. 2015. Hypothesis testing for differences in gaussian graphical models: Applications to brain connectivity. *arXiv preprint arXiv:1512.08643*.

Bingham, Ella, and Heikki Mannila. 2001. Random projection in dimensionality reduction: applications to image and text data. In *Sigkdd ickddm*.

Bonneel, Nicolas, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. 2014. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision* 51.

Cai, T., W. Liu, et al. 2011. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *JASA* 106(494):594–607.

Carlini, Nicholas, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Shuang Song, Abhradeep Thakurta, and Florian Tramer. 2020.

An attack on instahide: Is private learning possible with instance encoding? *arXiv preprint arXiv:2011.05315*.

Carlini, Nicholas, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} security symposium ({USENIX} security 19)*, 267–284.

Carroll, J Douglas, and Jih-Jie Chang. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika* 35(3).

Cohen, Nadav, Or Sharir, and Amnon Shashua. 2016. On the expressive power of deep learning: A tensor analysis. In *Conference on learning theory*, 698–728.

Cohen, Nadav, and Amnon Shashua. 2016. Convolutional rectifier networks as generalized tensor decompositions. In *International conference on machine learning*, 955–963.

Cornea, E., H. Zhu, et al. 2016. Regression models on Riemannian symmetric spaces. *JRSS-B*.

Cuturi, Marco. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* 26:2292–2300.

Danaher, P., P. Wang, et al. 2014. The joint graphical LASSO for inverse covariance estimation across multiple classes. *JRSS-B* 76(2):373–397.

Doshi-Velez, Finale, and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Fletcher, T.P., and S. Joshi. 2007. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing* 87(2):250–262.

Friedman, J., T. Hastie, et al. 2008. Sparse inverse covariance estimation with the graphical LASSO. *Biostatistics* 9(3):432–441.

FTC. 2021. California company settles ftc allegations it deceived consumers about use of facial recognition in photo storage app.

Geer, S.A. 2000. *Empirical processes in m-estimation*, vol. 6. Cambridge university press.

Ginart, A, M Guan, G Valiant, and J Zou. 2019. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*.

Golatkar, Aditya, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. 2021. Mixed-privacy forgetting in deep networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)*, 792–801.

Golatkar, Aditya, Alessandro Achille, and Stefano Soatto. 2020a. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 9304–9312.

———. 2020b. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *European conference on computer vision*, 383–398. Springer.

Hardle, Wolfgang, and Enno Mammen. 1993. Comparing nonparametric versus parametric regression fits. *The Annals of Statistics* 1926–1947.

Harshman, Richard A. 1970. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis.

Harvey, Jules., Adam. LaPlace. 2021. Exposing.ai.

Huang, Haiwen, Zhihan Li, Lulu Wang, Sishuo Chen, Bin Dong, and Xinyu Zhou. 2020. Feature space singularity for out-of-distribution detection. *arXiv preprint arXiv:2011.14654*.

Jayasumana, S., et al. 2013. Kernel methods on the Riemannian manifold of SPD matrices. In *Cvpr*.

Jiang, Ray, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. 2020. Wasserstein fair classification. In *Uncertainty in artificial intelligence*, 862–872. PMLR.

Jordan, M.I. 1998. *Learning in graphical models*, vol. 89. Springer Science & Business Media.

Kaye, Kate. 2022. The ftc's new enforcement weapon spells death for algorithms.

Khrulkov, Valentin, Oleksii Hrinchuk, and Ivan Oseledets. 2019. Generalized tensor models for recurrent neural networks. In *International conference on learning representations*.

Kim, Hyunwoo J, Nagesh Adluru, et al. 2014. Multivariate general linear models on Riemannian manifolds with applications to statistical analysis of diffusion weighted images. In *Cvpr*, 2705–2712.

Kline, Jeffery. 2019. Properties of the d-dimensional earth mover's problem. *Discrete Applied Mathematics* 265:128–141.

Koller, D., and N. Friedman. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Kwegyir-Aggrey, Kweku, Rebecca Santorella, and Sarah M. Brown. 2021. Everything is relative: Understanding fairness with optimal transport. 2102.10349.

Lauritzen, S.L. 1996. *Graphical models*. Clarendon Press.

Liu, H., J. Lafferty, et al. 2009. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *JMLR* 10:2295–2328.

Liu, Han, Fang Han, et al. 2012. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics* 40(4):2293–2326.

McArdle, J.J., and R.Q. Bell. 2000. An introduction to latent growth models for developmental data analysis.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54(6):1–35.

Mehta, Ronak, Rudrasis Chakraborty, Yunyang Xiong, and Vikas Singh. 2019a. Scaling recurrent models via orthogonal approximations in tensor trains. In *Proceedings of the ieee/cvf international conference on computer vision (iccv)*.

Mehta, Ronak, Hyunwoo Kim, Shulei Wang, Sterling Johnson, Ming Yuan, and Vikas Singh. 2019b. Localizing differentially evolving covariance structures via scan statistics. *Quarterly of Applied Math.* 77(2):357–398.

Mehta, Ronak, Sourav Pal, Vikas Singh, and Sathya N. Ravi. 2022. Deep unlearning via randomized conditionally independent hessians.

Mitchell, Tom M. 1997. *Machine learning*. New York: McGraw-Hill.

Nowozin, Sebastian, and Christoph H Lampert. 2011. *Structured learning and prediction in computer vision*, vol. 6. Now publishers Inc.

Oseledets, Ivan V. 2011. Tensor-train decomposition. *SIAM Journal on Scientific Computing* 33(5):2295–2317.

Qiu, H., F. Han, et al. 2015. Joint estimation of multiple graphical models from high dimensional time series. *JRSS-B*.

Ren, Jie, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. 2019. Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems* 32.

Roehrig, Charles S. 1988. Conditions for identification in nonparametric and parametric models. *Econometrica: Journal of the Econometric Society* 433–447.

Sekhari, Ayush, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. 2021. Remember what you want to forget: Algorithms for machine unlearning. 2103.03279.

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the ieee international conference on computer vision*, 618–626.

Solomon, Justin, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. 2015. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)* 34(4):1–11.

Städler, Nicolas, and Sach Mukherjee. 2012. Two-sample testing in high-dimensional models. arXiv:1210.4584.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems* 27.

Tansey, Wesley, Yixin Wang, David Blei, and Raul Rabadan. 2018. Black box fdr. In *International conference on machine learning*, 4867–4876. PMLR.

Ullman, J.B., and P.M. Bentler. 2003. *Structural equation modeling*. Wiley Online Library.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30.

Xie, Y., B.C. Vemuri, et al. 2010. Statistical analysis of tensor fields. In *Miccai*, 682–689. Springer.

Xue, L., and H.and others Zou. 2012. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics* 40(5):2541–2571.

Yang, S., Z. Lu, et al. 2015. Fused multiple graphical lasso. *SIAM J. Opt.* 25(2): 916–943.

Yang, Yinchong, Denis Krompass, and Volker Tresp. 2017. Tensor-train recurrent neural networks for video classification. In *Icml*.

Ye, Jieping, Ravi Janardan, and Qi Li. 2005. Two-dimensional linear discriminant analysis. In *Nips*, 1569–1576.

Yeung, Daniel S, Ian Cloete, Daming Shi, and Wing wY Ng. 2010. *Sensitivity analysis for neural networks*. Springer.

Yu, Xiyu, Tongliang Liu, Xinchao Wang, and Dacheng Tao. 2017. On compressing deep models by low rank and sparse decomposition. In *Cvpr*, 7370–7379.

Yuan, M. 2010. High dimensional inverse covariance matrix estimation via LP. *JMRL* 11:2261–2286.

Yurochkin, Mikhail, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin Solomon. 2019. Hierarchical optimal transport for document representation. 1906.10827.

Zhang, Qingchen, Laurence T Yang, Xingang Liu, Zhikui Chen, and Peng Li. 2017. A tucker deep computation model for mobile multimedia feature learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13(3s):39.

Zhang, Ye, and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Zhou, S., J. Lafferty, et al. 2010. Time varying undirected graphs. *ML* 80(2-3):295–319.