Try to be close instead of right: Consistent parameter manifolds generalize regularization theory

W. Evan Durno

January 10, 2018

Abstract

Regularization of statistical models is known to decrease variance and increase bias. It is an essential tool in this era of Big Data. Yet many regularization studies are largely empirical. Modern theory lacks generality, leaving many cases unexplained. This work brings generality to regularization theory through the study of parameters spaces as manifolds, requiring them to approach a limit which includes the *true* parameter value, hence attaining a form of consistency. It is shown that this assumption is already common in popular contexts. Using expected square distance as a measure of estimate error, it is proven that if being right requires estimation of too many parameters, then it is better to just be close.

For applied statisticians, this work has a simple lesson: build consistency into regularization schemes. For theoreticians, this work describes geometric problems in an important part of statistical theory.

Keywords: Regularization, Alternative hypothesis, Manifold, Geometry

1. Introduction

Mathematical models are an excellent proxy for reality because they are far simpler. They are simple enough that we use them to describe, understand, and predict. Some situations permit the definition of all variables and nearly perfect inference of their dependencies in a high-quality model, like the modelling of a swinging pendulum. Models can also be wrong. To be *less wrong*, one might define a model with parameters, moving parts, that allow a model to conform to data. Fitting a model to data requires taking several observations. Fitting a deterministic model will likely need as many observations as it does parameters. Random models tend to need far more.

While using parameters may be necessary to be less wrong, they ironically bring additional error. If parameters' estimates have substantial error, then models may still be uselessly wrong. In the fitting of parameterized models, this makes reduction of estimation error a necessary goal.

Applying statistical theory to solve real-world problems requires an understanding of how sample size effects estimation error. It is generally true that increasing sample sizes will decrease estimator variances but it is also widely observed that increasing the number of parameters also increases estimation error. Determining parameter relevance can sometimes be done. For example, the AIC statistic [1] describes how dissimilar an estimated model is from the unknown *true* model. For larger data sets, parameter definition is complicated, thereby necessitating feature design and selection.

The term *regularization* has come to refer to a variety of methods used to reduce estimation error. *Ridge Regression* [7] is a classic example, where parameters are estimated with a penalized optimization routine–punishing the estimation software for choosing large parameter values. Sometimes parameters are explicitly constrained to a lower-dimensional subspace through reparameterization [10]. The theory explaining these mechanisms tends to be domain-specific [3, 8, 2], and thereby lacks generality. Generalizing regularization theory is valuable both in guiding statisticians' intuition, and because empirical advances in statistics cannot push frontiers alone and tend to stand on theoretical foundations. For example, modern deep learning would not exist without thoroughly developing calculus [12]. This work generalizes regularization theory, unifying many popular and effective approaches by demonstrating the effect of already-popular methodologies and assumptions, albeit from a more abstract perspective. Results here pertain to regularization methods which work by constraining the parameter space to a manifold. Implicitly, this includes penalized likelihood methods, as illustrated in subsection 1.2.

Much of the philosophy of estimation is built around the assumed existence of a *true* model. With Result 2, we show that the discovery of any such model can be less important than finding one which is simply close enough. Specifically, if discovery of a true statistical model requires estimating parameters, then there exists a model which is wrong (biased) but with less estimation error than the larger true model. In the right situations, it is more pragmatic to try to be just a little wrong than try to be right. This highlights the feasibility of a strategy alternative to recovering a true model [8], where models need only be sufficiently flexible and not over-parameterized.

In this era of Big Data, better models tend to be larger and can't seem to be large enough. It seems entirely possible that some *true* models may actually be infinitely parameterized, effectively voiding any attempt at their recovery because every sample is finite. Fortunately that doesn't matter, because we need only be close enough.

1.1. Definitions

For a statistical model describing a simple, random sample, let the density or mass function of each observation be $f_X(X_i;\theta)$, where X_i is a random variable and θ is a model parameter. So the likelihood function is $\phi(X;\theta) = \prod_{i=1}^{n} f_X(X_i;\theta)$. Thus, the log likelihood function is $\log\phi(X;\theta) = \sum_{i=1}^{n} \log f_X(X_i,\theta)$.

The model's parameter θ may take on a *true* value, θ_1, but we are often only able to estimate a potentially biased $\hat{\theta}_0$. Let the true parameter space Θ_1 satisfy $\theta_1 \in \Theta_1$. Let the estimated parameter space Θ_0 satisfy $\hat{\theta}_0 \in \Theta_0$ and $\Theta_0 \subset \Theta_1$. For convenience, we require Θ_0 to be a closed set, with boundary set $\partial\Theta_0$. $\log\phi$ is assumed twice continuously differentiable (C^2) in θ.

Definition 1.

$\theta_1 := \arg\max_{\theta\in\Theta_1} \mathbb{E}\log f_X(X_1;\theta)$
$\theta_0 := \arg\max_{\theta\in\Theta_0} -\frac{1}{2}(\theta-\theta_1)^T\mathcal{I}_{\theta_1}(\theta-\theta_1)$
$\hat{\theta}_1 := \arg\max_{\theta\in\Theta_1} n^{-1}\log f_X(X;\theta)$
$\hat{\theta}_0 := \arg\max_{\theta\in\Theta_0} n^{-1}\log f_X(X;\theta)$
For matrix X, define $[X]_{a:b,c:d}$ be the submatrix of X containing rows a to b and columns c to d, inclusively counted.
Define $[X]_{:,c:d}$ as the submatrix of X containing columns c to d.
Similarly define $[x]_{a:b}$ as a sub-vector of entries a to b.

Note that expectation is always taken with $\theta = \theta_1$ because it is the *true* value. $\mathcal{I}_{\theta_1}$ is the Fisher Information Matrix, and is assumed positive definite ($\mathcal{I}_{\theta_1} > 0$). Expectations of the log likelihood and its first and second derivatives are assumed well-defined and finite.

$$[\mathcal{I}_{\theta_1}]_{jk} = -\mathbb{E}(\partial^2/(\partial\theta_j\partial\theta_k))\log f_X(X_1;\theta)|_{\theta=\theta_1}$$

Ultimately, this work compares estimation strategies through a measurement of estimation error, $\varepsilon(\theta) := \mathbb{E}||\theta_1 - \theta||_2^2$. It will be used to prove how manifold-constrained estimates $\hat{\theta}_0$ can have less error than their unconstrained counter-parts $\hat{\theta}_1$. So we ask, is $\varepsilon(\hat{\theta}_0) < \varepsilon(\hat{\theta}_1)$?

This work builds on the theory of alternative hypothesis testing to produce its results. It is assumed that Θ_1 is a manifold parameterized from $H_1 = \mathbb{R}^p$ by the C^1 diffeomorphism $\theta : H_1 \to \Theta_1$. Notice this requirement constrains the geometry of Θ_1. We also assume the existence of an $\eta_1 \in H_1$ corresponding to θ_1, satisfying $\theta_1 = \theta(\eta_1)$. Similarly, we assume the existence of $H_0 = \mathbb{R}^q \times \{[\eta_0]_{q+1}\} \times \cdots \times \{[\eta_0]_p\}$ where $q \leq p$, and we assume the existence of $\eta_0 \in H_0$ satisfying $\theta_0 = \theta(\eta_0)$. Notice that if $\eta \in H_0$, then $[\eta]_{q+1:p} = [\eta_0]_{q+1:p}$ is fixed per n. It is sometimes convenient to refer to Θ_1 as θ-space and to H_1 as η-space. Alternative hypothesis testing theory produces results in η-space for the estimates $\hat{\eta}_0 := \theta^{-1}(\hat{\theta}_0)$ and $\hat{\eta}_1 := \theta^{-1}(\hat{\theta}_1)$.

If $\partial\theta(\eta)$ parameterizes $\partial\Theta_0$, then it is constrained by the definition of $\theta(\eta)$. Particularly, because $\partial\Theta_0$ is defined by fixing $[\eta]_{q+1:p} = [\eta_0]_{q+1:p}$ constant per n, $\partial\theta(\eta)$ is constant over $[\eta]_{q+1:p}$, so $\partial\theta(\eta) = \partial\theta([\eta]_{1:q})$. Further, $\partial\Theta_0 = \{\theta(\eta) \ : \ [\eta]_{q+1:p} = [\eta_0]_{q+1:p}\}$, so $\partial\theta([\eta]_{1:q}) = \theta(\eta)|_{[\eta]_{q+1:p}=[\eta_0]_{q+1:p}}$. Also note that $\theta(\eta)|_{[\eta]_{1:q}=[\eta_0]_{1:q}}$ parameterizes the null space to $\partial\theta(\eta)$ near η_0 when ever $\theta(\eta)$ is fully differentiable.

This work must introduce the concept of a *consistent manifold* to build its argument. It will be shown that, while previously undefined, the concept is already in popular usage. An estimate $\hat{\theta}_1$ is consistent if it converges in measure to its true value, $\hat{\theta}_1 \to_{\mathbb{P}} \theta_1$. Because this work constrains the estimate $\hat{\theta}_0$ to a manifold Θ_0, it is possible that the manifold is *biased* $\theta_1 \notin \Theta_0$ thereby denying estimator consistency. To overcome this issue, we require Θ_0 to deform with the sample size n so that $\theta_1 \in \lim_{n\to\infty} \Theta_0$ and thereby describe Θ_0 as a *consistent manifold*. For example, Θ_0 can be biased $\theta_1 \notin \Theta_0$, and consistent $\theta_1 \in \lim_{n\to\infty} \Theta_0$. This results in $\lim_{n\to\infty} \theta_0 = \theta_1$ and thus $\lim_{n\to\infty} \eta_0 = \eta_1$. This work assumes there is an $s \in [0, 1/2)$ such that $\lim_{n\to\infty} n^s ||\theta_0 - \theta_1||_1 = 0$, thereby requiring $\sqrt{n}^{-1} ||\theta_0 - \theta_1||_1 \to 0$, but allowing $\sqrt{n}||\theta_0 - \theta_1||_1$ to model different kinds of bias. Further, this convergence requirement allows Θ_0 to converge discontinuously or always equate with its limit.

1.2. Consistent manifolds in popular usage

The popular *Lasso* estimate for least-squares regression is solved for via the following optimization problem.

$$\hat{\theta}_0 = \arg\min_{\beta\in\mathbb{R}^q} ||X\beta - Y||_2^2 + \lambda||\beta||_1$$

Similarly, the *glmnet* [6] regularized GLM estimator is able to produce solutions to the following optimization problem.

$$\hat{\theta}_0 = \arg\max_{\theta \in \mathbb{R}^q} \log \phi(Y, X; \theta) - \lambda ||\theta||_2^2$$

It can be shown that these optimization problems are equivalent to manifold-constrained estimates via the Lagrange Multiplier paradigm. First define the two following Lagrangians.

$$\Lambda'(\theta, \lambda) = \log \phi(Y, X; \theta) - \lambda ||\theta||_2^2$$

$$\Lambda(\theta, \lambda) = \log \phi(Y, X; \theta) + \lambda(||\hat{\theta}_0||_2^2 - ||\theta||_2^2)$$

Note that $\Lambda(\theta, \lambda)$ is the Lagrangian for the following constrained optimization program.

$$\arg\max_{\theta \in \mathbb{R}^q} \log \phi(Y, X; \theta) \ : \ ||\theta||_2^2 = ||\hat{\theta}_0||_2^2,$$

where the $A \ : \ B$ denotes A *such that* B. The existence of a solution to this program implies that it exists at a critical point $(\hat{\theta}_0, \lambda)$ of its Lagrangian, $\nabla_{\theta,\lambda} \Lambda(\theta, \lambda) = 0$, where $\nabla_x f$ is the gradient vector of f in x. A key insight to developing this critical point is that it is constrained by via $\nabla_\theta \Lambda' = \nabla_\theta \Lambda$. Further note that λ is not solved for, but given in its regularized usage. These constraints are used to construct a critical point, which is later proven optimal.

Assuming local differentiability, and $\nabla_\theta ||\theta||_2^2 \neq 0$, the regularized estimate $\hat{\theta}_0 = \arg\max_{\theta \in \mathbb{R}^q} \Lambda'(\theta, \lambda)$ occurs at its own critical point $\nabla_\theta \Lambda'(\theta, \lambda) = 0$, thereby implying $\lambda = (\nabla_\theta \log \phi(Y, X; \theta))/(\nabla_\theta ||\theta||_2^2)|_{\theta = \hat{\theta}_0}$. Hence $\hat{\theta}_0$ is constrained by the choice of regularization parameter λ.

Similarly, any critical point of $\Lambda(\theta, \lambda)$ must satisfy two requirements: (1) $\nabla_\theta \log \phi(Y, X; \theta) = \lambda \nabla_\theta ||\theta||_2^2$ and (2) $||\theta||_2^2 = ||\hat{\theta}_0||_2^2$. By the definition of $\hat{\theta}_0$, the first requirement is satisfied. Choosing $\theta = \hat{\theta}_0$ clearly satisfies the second requirement. Hence $(\hat{\theta}_0, \lambda)$ is a critical point for $\Lambda(\theta, \lambda)$. If $\hat{\theta}_0$ uniquely satisfies the requirements, then Lagrange Multiplier theory determines that it must be the solution to the constrained optimization program.

Uniqueness is implicitly given. Say there is some $\hat{\theta}'_0$ such that $\hat{\theta}'_0 \neq \hat{\theta}_0$, $||\hat{\theta}'_0||_2^2 = ||\hat{\theta}_0||_2^2$, and $\nabla_\theta \log \phi(Y, X; \theta)|_{\theta = \hat{\theta}'_0} = \lambda \nabla_\theta ||\theta||_2^2|_{\theta = \hat{\theta}'_0}$. Then $\hat{\theta}'_0$ is a

critical point, $\nabla_\theta \Lambda'(\theta, \lambda)|_{\theta=\hat{\theta}'_0} = 0$, but $\Lambda'(\hat{\theta}'_0, \lambda) < \Lambda'(\hat{\theta}_0, \lambda)$ because it is not the regularized estimate. This implies sub-optimality of $\hat{\theta}'_0$ as follows.

$$\Lambda'(\hat{\theta}'_0, \lambda) < \Lambda'(\hat{\theta}_0, \lambda)$$
$$\Leftrightarrow \log \phi(Y, X; \hat{\theta}'_0) - \lambda ||\hat{\theta}'_0||_2^2 = \log \phi(Y, X; \hat{\theta}'_0) - \lambda ||\hat{\theta}_0||_2^2 < \log \phi(Y, X; \hat{\theta}_0) - \lambda ||\hat{\theta}_0||_2^2$$
$$\Leftrightarrow \log \phi(Y, X; \hat{\theta}'_0) < \log \phi(Y, X; \hat{\theta}_0)$$

Hence the L_2 regularized estimate for a GLM is also the solution to a manifold-constrained estimate. Particularly $\Theta_0 = \{\theta \in \mathbb{R}^q \ : \ ||\theta||_2^2 = ||\hat{\theta}_0||_2^2\}$. Using a slightly longer argument, it possible to derive the same result for a variety of voluminous manifolds, including the following constrained optimization program where $||\hat{\theta}_0||_2^2 \leq d$.

$$\hat{\theta}_0 = \arg\max_{\theta \in \mathbb{R}^q} \log \phi(Y, X; \theta) \ : \ ||\theta||_2^2 \leq d$$

To generalize the result to manifolds with merely almost-everywhere (*a.e.*) differentiable boundaries, realize that if the boundary is composed of *a.e.*-differentiable facets, then we recognize that each facet has a different probability of containing $\hat{\theta}_0$. So regularization becomes equivalent to several cases of constrained optimization programs, each solvable with Lagrange Multipliers, and each with its own probability of occurring. For example, *Lasso*-regularized $||\theta||_1$ estimates fall into a manifold with a not-completely, yet *a.e.*-differential boundary. In this case, individual facets and ridges are still parameterizable as Lagrange Multiplier constraints. A interesting degenerate case is where $\hat{\theta}_0$ falls to a corner of the *Lasso* polytope Θ_0 with high-probability, in which case the optimization program constrains all candidate solutions to a single point.

Having described these manifolds, it is now essential to argue their consistency to finish our claim that consistent manifolds are already in popular usage. The key insight follows.

$$\arg\max_{\theta \in \mathbb{R}^q} \log \phi(Y, X; \theta) - \lambda ||\theta||_2^2 = \arg\max_{\theta \in \mathbb{R}^q} n^{-1} \log \phi(Y, X; \theta) - n^{-1} \lambda ||\theta||_2^2$$

Because $\lim_{n \to \infty} n^{-1} \log \phi(Y, X; \theta_1) = \mathbb{E} \log f_{Y,X}(Y_1, X_1; \theta_1)$ *a.s.* (almost surely) and $\lim_{n \to \infty} n^{-1} \lambda ||\theta_1||_2^2 = 0$, hence approaching an unregularized, correct estimate, the associated manifold Θ_0 must approach Θ_1 which contains θ_1, thereby attaining consistency.

This work proves how dimensional reduction of parameter spaces can reduce estimator error. Notice that when shrinkage estimators are equivalent to estimation on a biased, ball-shaped manifold, estimates can be constrained to the ball's boundary. When locally differentiable, the boundary is a manifold with lower dimension than the interior. So while error may be gained in bias, constraint to a low-dimensional boundary will reduce error via less variance.

2. η-space results

Given parameter manifold consistency and regularity assumptions, the distribution of a null-constrained estimate under the alternative hypothesis can be derived. Null hypotheses, which specify exact parameter values (like $H_0 : [\eta]_{q+1:p} = 0$), effectively constrain the estimate $\hat{\eta}_0$ to a dimensionally-reduced manifold. While the shape of this manifold H_0 may not be equivalent to the regularization manifold θ_0 (which is often a closed, bounded ball), the existence of a diffeomorphism $\theta(\eta)$ between H_0 and Θ_0 permits η-space theory to be generalized to θ-space. This construction provides nearly immediate results for regularization theory.

The work in this section draws a great deal of inspiration from Davidson and Lever [4]. While their results were constrained to η-space thereby avoiding a manifold-based interpretation of estimation, Davidson and Lever did assume that $||\eta_0 - \eta_1|| \to 0$ as $n \to \infty$, which is implicit in our assumption of parameter manifold consistency. A superficial differentiation is that their consistency assumption is achieved via $\eta_1 \to \eta_0$ as $n \to \infty$, thereby requiring true parameter values to change with the number of samples. As demonstrated in subsection 1.2, manifolds implicit in popular regularization schemes tend to reduce bias with sample size, so our work assumes $\eta_0 \to \eta_1$ as $n \to \infty$.

2.1. $\hat{\eta}_0$ asymptotics

Definition 2.

For function $f \in \mathbb{R}^a$ and differentiable in $x \in \mathbb{R}^b$, the Jacobian is $J_{f,x} := \left[\frac{\partial f_i}{\partial x_j}(x)\right] \in \mathbb{R}^{a \times b}$.

For $\theta(\eta)$ evaluated at $\theta^{-1}(\theta)$, define $J_\theta := J_{\theta,\theta^{-1}(\theta)}$.

Let N_p denote an $N_p(0, I_p)$-distributed random variable.

$\mathcal{I}_{\eta_1} := J_{\theta_1}^T \mathcal{I}_{\theta_1} J_{\theta_1}$

$\mathcal{I}_{\eta_0} := [\mathcal{I}_{\eta_1}]_{1:q,1:q} = [J_{\theta_1}]_{:,1:q}^T \mathcal{I}_{\theta_1} [J_{\theta_1}]_{:,1:q}$

Note that $\mathcal{I}_{\eta_0} > 0$.

Lemma 1.

$$\sqrt{n}[\hat{\eta}_0 - \eta_1]_{1:q} + O(\sqrt{n}[\eta_0 - \eta_1]_{q+1:p}) \to_d \mathcal{I}_{\eta_0}^{-1/2} N_q \text{ as } n \to \infty$$

Where $A \to_d B$ denotes convergence in distribution, and $[\hat{\eta}_0]_{1:q}$ is estimated, while $[\hat{\eta}_0]_{q+1:p} = [\eta_0]_{q+1:p}$ is fixed per n.

Proof. of Lemma 1

$\hat{\eta}_0 = \arg\max_{\eta \in H_0} \log \phi(X; \theta(\eta)) = \arg\max_{\eta \in H_0} \sqrt{n}^{-1} \log \phi(X; \theta(\eta))$

$= \arg\max_{\eta \in H_0} \sqrt{n}^{-1} \log \phi(X; \theta(\eta_1)) + (\eta - \eta_1)^T \nabla_\eta \sqrt{n}^{-1} \log \phi(X; \theta(\eta_1))$

$+\frac{1}{2}(\eta - \eta_1)^T \nabla_\eta^2 \sqrt{n}^{-1} \log \phi(X; \theta(\eta_1))(\eta - \eta_1) + O(\sqrt{n}^{-1}||\eta - \eta_1||_2^2)$

Remark: $\nabla_\eta \log \phi(X; \theta(\eta_1)) = \nabla_\eta \log \phi(X; \theta(\eta))|_{\eta=\eta_1}$ is a gradient vector.

Remark: $\nabla_\eta^2 \log \phi(X; \theta(\eta_1))$ is a Hessian.

$= \arg\max_{\eta \in H_0} 0 + (\eta - \eta_1)^T \nabla_\eta \sqrt{n}^{-1} \log \phi(X; \theta(\eta_1))$

$+\frac{1}{2}(\eta - \eta_1)^T \nabla_\eta^2 \sqrt{n}^{-1} \log \phi(X; \theta(\eta_1))(\eta - \eta_1) + O(\sqrt{n}^{-1}||\eta - \eta_1||_2^2)$

$= \arg\max_{\eta \in H_0} (\eta - \eta_1)^T J_{\theta_1}^T \nabla_\theta \sqrt{n}^{-1} \log \phi(X; \theta_1)$

$+\frac{1}{2}(\eta - \eta_1)^T J_{\theta_1}^T \nabla_\theta^2 \sqrt{n}^{-1} \log \phi(X; \theta_1) J_{\theta_1}(\eta - \eta_1) + O(\sqrt{n}^{-1}||\eta - \eta_1||_2^2)$

$= \arg\max_{\eta \in H_0} (\eta - \eta_1)^T J_{\theta_1}^T \nabla_\theta \sqrt{n}^{-1} \sum_{i=1}^n \log f_X(X_i; \theta_1)$

$+\frac{1}{2}(\eta - \eta_1)^T J_{\theta_1}^T \nabla_\theta^2 n^{-1} \sum_{i=1}^n \log f_X(X_i; \theta_1) J_{\theta_1} \sqrt{n}(\eta - \eta_1)$

$+O(\sqrt{n}^{-1}||\eta - \eta_1||_2^2)$

Remark: Find the critical point for $[\eta]_{1:q}$.

$0 = \nabla_{[\eta]_{1:q}} \Big((\eta - \eta_1)^T J_{\theta_1}^T \nabla_\theta \sqrt{n}^{-1} \sum_{i=1}^n \log f_X(X_i; \theta_1)$

$+\frac{1}{2}(\eta - \eta_1)^T J_{\theta_1}^T \nabla_\theta^2 n^{-1} \sum_{i=1}^n \log f_X(X_i; \theta_1) J_{\theta_1} \sqrt{n}(\eta - \eta_1)$

$+O(\sqrt{n}^{-1}||\eta - \eta_1||_2^2) \Big)$

$= [J_{\theta_1}]_{:,1:q}^T \nabla_\theta \sqrt{n}^{-1} \sum_{i=1}^n \log f_X(X_i; \theta_1)$

$+[J_{\theta_1}]_{:,1:q}^T \nabla_\theta^2 n^{-1} \sum_{i=1}^n \log f_X(X_i; \theta_1)[J_{\theta_1}]_{:,1:q} \sqrt{n}[\eta - \eta_1]_{1:q}$

$+[J_{\theta_1}]_{:,1:q}^T \nabla_\theta^2 n^{-1} \sum_{i=1}^n \log f_X(X_i; \theta_1)[J_{\theta_1}]_{:,q+1:p} \sqrt{n}[\eta - \eta_1]_{q+1:p})$

$+O(\sqrt{n}^{-1}(\eta - \eta_1))$

Remark: Apply the Central Limit Theorem and the Strong Law of Large Numbers to get convergence in distribution.

$\to_d [J_{\theta_1}]_{:,1:q}^T \mathcal{I}_{\theta_1}^{1/2} N_p - [J_{\theta_1}]_{:,1:q}^T \mathcal{I}_{\theta_1} [J_{\theta_1}]_{:,1:q} \sqrt{n}[\eta - \eta_1]_{1:q}$

$-[J_{\theta_1}]_{:,1:q}^T \mathcal{I}_{\theta_1} [J_{\theta_1}]_{:,q+1:p} \sqrt{n}[\eta - \eta_1]_{q+1:p})$

$=_d \mathcal{I}_{\eta_0}^{1/2} N_q - \mathcal{I}_{\eta_0} \sqrt{n}[\eta - \eta_1]_{1:q}$

$-[J_{\theta_1}]_{:,1:q}^T \mathcal{I}_{\theta_1} [J_{\theta_1}]_{:,q+1:p} \sqrt{n}[\eta - \eta_1]_{q+1:p}$

Remark: $\mathcal{I}_{\eta_0} > 0 \Rightarrow \exists \mathcal{I}_{\eta_0}^{-1/2}$, but $\mathcal{I}_{\eta_0}^{-1/2} \neq ([J_{\theta_1}]_{:,1:q}^T \mathcal{I}_{\theta_1}^{1/2})^{-1}$ which does not exist.
$\Rightarrow \sqrt{n}([\eta]_{1:q} - [\eta_1]_{1:q}) + O(\sqrt{n}[\eta - \eta_1]_{q+1:p}) =_d \mathcal{I}_{\eta_0}^{-1/2} N_q$
Large n makes $\log \phi$ convex, so $[\hat{\eta}_0]_{1:q} = [\eta]_{1:q}$ uniquely.

□

Lemma 2.
$\hat{\eta}_0$ is consistent.

Proof. of Lemma 2
By Lemma 1, $\sqrt{n}[\hat{\eta}_0 - \eta_1]_{1:q} \to_d \mathcal{I}_{\eta_0}^{-1/2} N_q + O(\sqrt{n}[\eta_0 - \eta_1]_{q+1:p})$
$\Rightarrow [\hat{\eta}_0 - \eta_1]_{1:q} \to_d \sqrt{n}^{-1} \mathcal{I}_{\eta_0}^{-1/2} N_q + O([\eta_0 - \eta_1]_{q+1:p})$
$\Rightarrow [\hat{\eta}_0 - \eta_1]_{1:q} \to_d \Rightarrow [\hat{\eta}_0 - \eta_1]_{1:q} \to_{\mathbb{P}} 0$

□

Lemma 3.

$$\sqrt{n}(\hat{\eta}_1 - \eta_1) \to_d \mathcal{I}_{\eta_1}^{-1/2} N_p \text{ as } n \to \infty$$

Proof. of Lemma 3
$\hat{\eta}_1$ is a MLE (maximum likelihood estimate), bijectively transformed through $\theta(\eta)$ and thereby remains an MLE by functional invariance. The covariance matrix is calculable via a Taylor series approximation.

□

3. $\boldsymbol{\theta}$-space results

Definition 3.
$\mathcal{I}_{\theta_0} := [J_{\theta_0}]_{:,1:q} \mathcal{I}_{\eta_0} [J_{\theta_0}]_{:,1:q}^T$
$\mathcal{I}'_{\theta_0} := [J_{\theta_0}]_{:,1:q} \mathcal{I}_{\eta_0}^{-1} [J_{\theta_0}]_{:,1:q}^T$
$\mathcal{I}''_{\theta_0} := [J_{\theta_0}]_{:,1:q} \mathcal{I}_{\eta_0}^{-1/2}$

Because $\text{Rank}([J_{\theta_0}]_{:,1:q}^T) \leq q < p$, $\mathcal{I}'_{\theta_0}$ can be singular.

3.1. $\hat{\theta}_0$ asymptotics

Lemma 4.
(i) $\hat{\theta}_0 \to_{\mathbb{P}} \theta_0$ as $n \to \infty$
(ii) If $\theta_1 \notin \Theta_0$ and n large, then $\sqrt{n}(\hat{\theta}_0 - \theta_0) + O(\sqrt{n}(\theta_0 - \theta_1)) \sim N_p(0, \mathcal{I}'_{\theta_0})$.

Proof. of Lemma 4

CASE $\theta_1 \in \Theta_0$:
$\hat{\theta}_0 = \hat{\theta}_1$ which is an unconstrained MLE,
so $\hat{\theta}_0 \to \theta_1$ and $\theta_0 = \theta_1$.
CASE $\theta_1 \notin \Theta_0$:
By MLE functional invariance, $\hat{\theta}_0 = \theta(\hat{\eta}_0)$.
Hence $\hat{\theta}_0 \to_{\mathbb{P}} \theta_0 = \theta^{-1}(\eta_0)$ by Lemma 2.
Remark: Postulate (i) is proven.
By Taylor Series approximation, $\hat{\theta}_0 = \theta_0 + J_{\theta_0}(\hat{\eta}_0 - \eta_0) + e_\eta(\hat{\eta}_0 - \eta_0)$,
where e_η is bounded near $\eta = \eta_0$ and $\lim_{n\to\infty} e_\eta = 0$.
$\Rightarrow \sqrt{n}(\hat{\theta}_0 - \theta_0) = \sqrt{n}(J_{\theta_0} + e_\eta)(\hat{\eta}_0 - \eta_0)$
$= \sqrt{n}[J_{\theta_0} + e_\eta]_{:,1:q}[\hat{\eta}_0 - \eta_0]_{1:q} + \sqrt{n}[J_{\theta_0} + e_\eta]_{:,q+1:p}[\hat{\eta}_0 - \eta_0]_{q+1:p}$
Remark: $[\hat{\eta}_0 - \eta_0]_{q+1:p} = 0$ by definition.
$= \sqrt{n}[J_{\theta_0} + e_\eta]_{:,1:q}[\hat{\eta}_0 - \eta_0]_{1:q} + 0$
Remark: Apply Lemma 1.
$\to_d [J_{\theta_0}]_{:,1:q}\mathcal{I}_{\eta_0}^{-1/2}N_q + O(\sqrt{n}[\eta_0 - \eta_1]_{q+1:p})$
Remark: $O([\eta_0 - \eta_1]_{q+1:p}) \approx O(\theta_0 - \theta_1)$ for n large. See the proof of Lemma 1 for the exact form of $O([\eta_0 - \eta_1]_{q+1:p})$.
$\Rightarrow \sqrt{n}(\hat{\theta}_0 - \theta_0) + O(\sqrt{n}(\theta_0 - \theta_1)) \to_d [J_{\theta_0}]_{:,1:q}\mathcal{I}_{\eta_0}^{-1/2}N_q =_d \mathcal{I}''_{\theta_0}N_p$
Remark: Postulate (ii) is proven.

□

3.2. Unbiased regularization always reduces error

Result 1. *If $q < p$, $\theta_0 = \theta_1$ and n large, then $\varepsilon(\hat{\theta}_0) < \varepsilon(\hat{\theta}_1)$.*

Proof. of Result 1

$\varepsilon(\hat{\theta}_0) < \varepsilon(\hat{\theta}_1) \Leftrightarrow \mathbb{E}||\hat{\theta}_0 - \theta_1||^2 < \mathbb{E}||\hat{\theta}_1 - \theta_1||^2 \Leftrightarrow \mathbb{E}||\hat{\theta}_0 - \theta_0||^2 < \mathbb{E}||\hat{\theta}_1 - \theta_1||^2$
Remark: Applied $\theta_0 = \theta_1$.
$\Leftrightarrow \sum_{j=1}^{p} \mathrm{Var}([\hat{\theta}_0]_j) < \sum_{j=1}^{p} \mathrm{Var}([\hat{\theta}_1]_j)$
Remark: Apply Lemma 4 (ii).
$\Leftrightarrow \mathrm{tr}([J_{\theta_0}]_{:,1:q}\mathcal{I}_{\eta_0}^{-1}[J_{\theta_0}]_{:,1:q}^T) = \mathrm{tr}\left(J_{\theta_0}\begin{bmatrix}\mathcal{I}_{\eta_0}^{-1} & 0\\ 0 & 0\end{bmatrix}J_{\theta_0}^T\right) < \mathrm{tr}(J_{\theta_1}\mathcal{I}_{\eta_1}^{-1}J_{\theta_1}^T)$
Remark: Apply $\theta(\eta)$ continuity: $\epsilon_n > 0$ and $\lim_{n\to\infty}\epsilon_n = 0$.
$\Leftrightarrow \mathrm{tr}\left(J_{\theta_1}\begin{bmatrix}\mathcal{I}_{\eta_0}^{-1} & 0\\ 0 & 0\end{bmatrix}J_{\theta_1}^T\right) + \epsilon_n < \mathrm{tr}(J_{\theta_1}\mathcal{I}_{\eta_1}^{-1}J_{\theta_1}^T)$
$\Leftrightarrow \epsilon_n < \mathrm{tr}(J_{\theta_1}\mathcal{I}_{\eta_1}^{-1}J_{\theta_1}^T) - \mathrm{tr}\left(J_{\theta_1}\begin{bmatrix}\mathcal{I}_{\eta_0}^{-1} & 0\\ 0 & 0\end{bmatrix}J_{\theta_1}^T\right) = \mathrm{tr}\left(J_{\theta_1}\left(\mathcal{I}_{\eta_1}^{-1} - \begin{bmatrix}\mathcal{I}_{\eta_0}^{-1} & 0\\ 0 & 0\end{bmatrix}\right)J_{\theta_1}^T\right)$
Remark: $P > 0 \Rightarrow \exists A \ : \ P = AA^T \Rightarrow JPJ^T = (JA)(AJ)^T$.

Remark: Logical equivalency follows when n is sufficiently large.
$\Leftrightarrow \operatorname{tr}\left(\begin{bmatrix}\mathcal{I}_{\eta_0}^{-1} & 0\\ 0 & 0\end{bmatrix}\right) < \operatorname{tr}(\mathcal{I}_{\eta_1}^{-1})$

$$= \operatorname{tr}\left(\begin{bmatrix}\mathcal{I}_{\eta_0}^{-1} + \mathcal{I}_{\eta_0}^{-1}[\mathcal{I}_{\eta_1}]_{1:q,q+1:p}K[\mathcal{I}_{\eta_1}]_{1:q,q+1:p}^T\mathcal{I}_{\eta_0}^{-1} & -\mathcal{I}_{\eta_0}^{-1}[\mathcal{I}_{\eta_1}]_{1:q,q+1:p}K\\ -K[\mathcal{I}_{\eta_1}]_{1:q,q+1:p}^T\mathcal{I}_{\eta_0}^{-1} & K\end{bmatrix}\right)$$

Remark: $K = ([\mathcal{I}_{\eta_1}]_{q+1:p,q+1:p} - [\mathcal{I}_{\eta_1}]_{1:q,q+1:p}^T\mathcal{I}_{\eta_0}^{-1}[\mathcal{I}_{\eta_1}]_{1:q,q+1:p})^{-1}$
$= \operatorname{tr}(\mathcal{I}_{\eta_0}^{-1}) + \operatorname{tr}(\mathcal{I}_{\eta_0}^{-1}[\mathcal{I}_{\eta_1}]_{1:q,q+1:p}K[\mathcal{I}_{\eta_1}]_{1:q,q+1:p}^T\mathcal{I}_{\eta_0}^{-1}) + \operatorname{tr}(K)$
Remark: $K > 0 \Rightarrow \operatorname{tr}(K) > 0$
$> \operatorname{tr}\left(\begin{bmatrix}\mathcal{I}_{\eta_0}^{-1} & 0\\ 0 & 0\end{bmatrix}\right)$

□

3.3. Biased regularization

Definition 4.
$\delta := \theta_0 - \theta_1$

We interpret δ as bias.

Lemma 5. *For n sufficiently large, $\mathbb{E}||\hat{\theta}_0 - \theta_0||_2^2 < \varepsilon(\hat{\theta}_1)$.*

Lemma 6. *$\varepsilon(\hat{\theta}_0) \to \mathbb{E}||\hat{\theta}_0 - \theta_0||_2^2 + ||\delta||_2^2$ as $n \to \infty$.*

Result 2. *If $q < p$ and $\theta_0 \neq \theta_1$, then there exists n sufficiently large that Θ_0 is both biased and $\varepsilon(\hat{\theta}_0) < \varepsilon(\hat{\theta}_1)$.*

Proof. of Result 2
By Lemma 5, $\mathbb{E}||\hat{\theta}_0 - \theta_0||_2^2 < \varepsilon(\hat{\theta}_1)$ is given for n large.
Θ_0 consistent implies $||\delta||_2^2 \to 0$ as $n \to \infty$, so for sufficiently large n,
$\varepsilon(\hat{\theta}_0) \to \mathbb{E}||\hat{\theta}_0 - \theta_0||_2^2 + ||\delta||_2^2 < \varepsilon(\hat{\theta}_1)$ by Lemma 6.

□

Result 2 proves that reducing a model's parameter dimension, even when incurring bias, can reduce the overall estimation error. For a neighbourhood of θ_1 it is better to be wrong.

Result 3. *If $\sqrt{n}||\delta||_1 \to \infty$ as $n \to \infty$, then $\mathbb{P}(\hat{\theta}_0 \in \partial\Theta_0) \to 1$.*

Proof. of Result 3

Observe that $\mathbb{P}(\hat{\theta}_0 \in \partial\Theta_0) = \mathbb{P}(\hat{\theta}_1 \notin \Theta_0^o)$,
where Θ_0^o is the ineterior of Θ_0. For large n, $\sqrt{n}(\hat{\theta}_1 - \theta_1) \sim N_p(0, \mathcal{I}_{\theta_1}^{-1})$,
because $\hat{\theta}_1$ is an unconstrained MLE. So $\mathbb{P}(\hat{\theta}_1 \in \Theta_0^o)$
$\leq \mathbb{P}(n(\hat{\theta}_1 - \theta_1)^T \mathcal{I}_{\theta_1}(\hat{\theta}_1 - \theta_1) \geq n(\theta_0 - \theta_1)^T \mathcal{I}_{\theta_1}(\theta_0 - \theta_1))$
$\leq \mathbb{P}(n(\hat{\theta}_1 - \theta_1)^T \mathcal{I}_{\theta_1}(\hat{\theta}_1 - \theta_1) \geq (h^T \sqrt{n}(\theta_0 - \theta_1))^2)$
$= \mathbb{P}(\sqrt{X_p^2} \geq h^T \sqrt{n}(\theta_0 - \theta_1)) \to 0$ as $n \to \infty$,
for $X_p^2 \sim \chi_p^2$ and $h \in \mathbb{R}_{>0}^p$ sufficiently small.

□

This demonstrates that for sufficiently biased parameter manifolds, estimates fall into $\partial\Theta_0$. We know from Results 1 and 2 that dimensional reduction of sufficiently unbiased parameter manifolds decreases error. Because $\partial\Theta_0$ has a lower dimension than Θ_0, bias incurs a dimensional reduction by Result 3. Thus a small amount of bias can decrease error.

3.4. Constructing $\theta(\eta)$

In attempting to use the theory developed here, and after defining Θ_1 and Θ_0, the statistician will need to confirm the existence of $\theta(\eta)$. Instead of burdening the applied statistician with discovering diffeomorphisms, Result 4 only requires the statistician to parameterize $\partial\Theta_0$ with a continuously differentiable $\partial\theta'(\eta)$. The result guarantees existence of $\theta(\eta)$ in an open subset set of Θ_1. If $\partial\theta'(\eta)$ is continuously differentiable at η_1, then the result is perfectly feasible for large sample sizes, because $\hat{\theta}_0 \to_{\mathbb{P}} \theta_1$ as $n \to \infty$. One might attempt to define a generally existing $\theta(\eta)$, but the task will not be necessary in most situations.

Result 4. *If $\partial\theta' : H_1 \to \theta_1$ parameterizes $\partial\Theta_0$ and $rank(J_{\partial\theta',\eta}) = q' \leq q \leq p$ constantly in U open $\subset H_1$, then there exists implicit diffeomorphism $\theta : U \to V$ open $\subset \Theta_1$ and boundary parameterization $\partial\theta(\eta) = \theta(\eta)|_{[\eta]_{q'+1:p}=0}$.*

Proof. of Result 4

By the Constant Rank Theorem, there exists diffeomorphisms $\psi : V \to V$ and $\phi : U \to U$ such that $(\psi \circ \partial\theta' \circ \phi^{-1})(\eta) = \eta|_{[\eta]_{q'+1:p}=0}$ for $\eta \in U$.
Since ψ is a diffeomorphism, ψ^{-1} is well defined.
Define $\theta(x) = \psi^{-1}\left((\psi \circ \partial\theta' \circ \phi^{-1})(x) + x|_{[x]_{1:q'}=0}\right)$,
where there exists $u \in U, v \in V$ for $[x]_{1:q'} = [u]_{1:q'}, [x]_{q'+1:p} = [v]_{q'+1:p}$.
Remark: Apply the inverse function theorem.

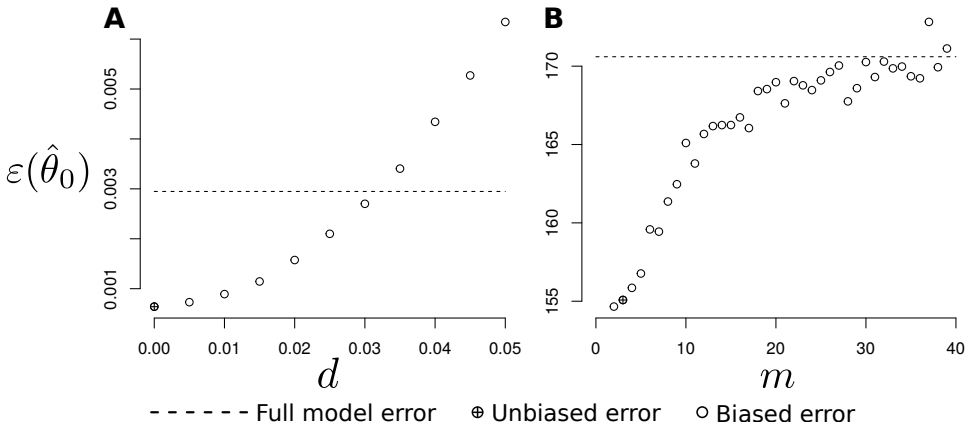

Figure 1: (A) Linear constraint between μ and σ^2 as regularization, (B) Factor Analysis Model as regularization

Since ψ is a diffeomorphism ψ^{-1} is continuously differentiable.
By the Inverse Function Theorem and chain rule,
$J^{-1}_{\theta,x} = J^{-1}_{\psi,x}\left(\begin{bmatrix} I_{q'} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & I_{p-q'} \end{bmatrix}\right)$.
Take $\partial\theta(x) = (\partial\theta' \circ \phi^{-1})(x)$.

□

4. Examples

4.1. Simple example

Let $\Theta_1 = \mathbb{R} \times \mathbb{R}_{>0}$ and $\theta_1 = (\mu_1, \sigma_1) = (1, 1)$. Data are $N_1(\mu_1, \sigma_1^2)$ distributed. Let Θ_0 be the line $(1 + t + d, 1 + t - d)$ in $\mathbb{R}^2$ parameterized by $t \in \mathbb{R}$, where d is fixed and models bias. Hence constrained parameter manifold is a line through $\mathbb{R}^2$. The relationship between estimation error $\varepsilon(\hat{\theta}_0)$ and bias d is illustrated in Figure 1 A. To produce the figure, $\varepsilon(\hat{\theta}_0)$ calculated as the average of 1000 independently estimated $\hat{\theta}_0$ values, each estimated with 1000 samples. Sample sizes must be large to demonstrate the difference between manifolds, because this is a low-dimensional parameter space. Nonetheless, the theory developed in this work stands true for models both large and small.

4.2. Factor analysis example

The Factor Analysis Model (FAM) constrains multivariate Gaussian covariance as follows.

$$X = LF + \Psi E, L \in \mathbb{R}^{d\times m}, \Psi \in \text{diag}(\mathbb{R}^{d\times d}_{>0}), F \sim N_m(0, I_m), E \sim N_d(0, I_d)$$

This causes $\text{Var}(X) = LL^T + \Psi^2$. In this example, Θ_1 is the set of all covariance matrices in $\mathbb{R}^{d\times d}$, and Θ_0 is the set of all FAM covariance matrices. θ_1 corresponds to an FAM with $m = 3$ and $d = 50$, causing some parameter manifolds Θ_0 to be biased and others not. Only 100 samples are drawn from the θ_1 distribution to fit each model. Every event of estimation is statistically independent. In Figure 1 B, simulations illustrate the relationship between $\varepsilon(\hat{\theta}_0)$ and m. Each $\varepsilon(\hat{\theta}_0)$ is estimated per m by averaging 100 trials. FAM estimates were computed using *factanal* in *R* version 3.3.2 [11], using *varimax* rotation.

Define $\Theta_0 = \{C \in \Theta_1 \ : \ \exists(L, \Psi) \in H_0 \text{ and } C = LL^T + \Psi^2\}$. With both Θ_0 and Θ_1 rigorously defined, application also requires proof of $\theta(\eta)$ existence. Instead of attempting to discover a generally existing $\theta(\eta)$, we will apply Result 4. Defining $\partial\theta'(L, \Psi) = LL^T + \Psi^2$, we can see $J_{\partial\theta',(L,\Psi)}$ is well defined and thus a $\theta(\eta)$ exists at least locally. Note that $\partial\theta'(\eta)$ is surjective and that bijectivity is not required by Result 4.

The last inexplicitly satisfied assumption is that of manifold consistency, because we have not defined Θ_0 as a function of n. Fortunately, we do know that increasing m to d with n would eventually produce an estimate for any covariance matrix Σ, including the non-unique $\Sigma = LL^T + 0$. Hence taking $m = m(n)$ for any increasing $m(n)$ satisfying $\lim_{n\to\infty} m(n) = d$ is sufficient. The arbitrary nature of selecting $m(n)$ highlights how mathematical generality is not the same as practical generality. The theory developed here only works because $||\theta_1 - \theta_0||$ is small. So in place of selecting $m(n)$, it would be more effective have a good reason to choose an FAM over other covariance constraints.

Demonstrating FAM as effective regularization is particularly relevant to this work's inspiring piece (see Chapter 3 of Durno [5] for more information).

5. Discussion

The choice of defining estimation error as $\varepsilon(\theta) = \mathbb{E}||\theta - \theta_1||_2^2$ is both effective and theoretically convenient. It is effective because the distance between an estimate and its true value is relevant to many statistical concepts. For example, sufficiently reducing estimation error must increase statistical power

for Neyman-Pearson Lemma [9]-type hypothesis tests. If we further make the Bayesian-style assumption of the population distributing amongst the null and alternative hypotheses ($\mathbb{P}(\theta = \theta_j | X_i)$ is well defined for $j \in \{1, 2\}$), then Naive Bayes classification is similar enough to a likelihood ratio test that improving statistical power becomes equivalent to increasing precision per fixed recall. We want our estimates to be close to their true values.

Showing that penalized likelihood regularization is equivalent to manifold-constrained estimation illustrates its similarity to dimensional reduction through Result 3. Particularly, if the implicit parameter manifold is sufficiently biased, then the estimate is constrained to the manifold's boundary for large sample sizes. Because the manifold's boundary is a lower-dimensional structure than the manifold's interior space, the estimate's dimension has been reduced. In turn, Result 2 shows that this biased dimensional reduction can still result in an overall reduction in estimation error.

Consistency of parameter manifolds is satisfied sometimes less explicitly than shown in subsection 1.2. For example, even if a parameter space is biased and not explicitly defined to deform over sample size, one may assume without any loss of mathematical generality that the parameter space is implicitly consistent, save any topological constraints. Of course, there is a loss of practical generality. The entire reason why the manifold consistency assumption is useful is because it ensures manifold bias can be sufficiently small for proof purposes. A practical solution is to simply define regularization methods with built-in consistency.

All theory in this work could be generalized to a case where manifolds are not assumed consistent, but bias is small. Knowing *how small is small enough* would be particularly valuable if manifolds could be very biased. When there is substantial room for error, it is hard to be wrong. If it is hard to be wrong, then any sufficiently flexible model can learn any system. So the description and separation of error into bias and variance is a worthwhile endevour. A potentially useful lead is that Lemma 4 can be proven without Lemma 1, but instead as the asymptotically unique solution to a linearly constrained quadratic program. It has a closed form solution, and much of the proof can keep variance and bias separate.

Asymptotic theory is a unifying tool, bringing general results to a wide variety of models. This work is no exception, but ironically regularization was initially designed for small-sample size applications. In this era of Big Data, there is practical room for nuance in this interpretation. If a true model can be estimated with a finite-dimensional parameter manifold, then there

are sufficiently large sample sizes where the bias incurred by regularization becomes impractical. If not, then regularization may always be valuable. Of course, this dichotomy might also be generalized. It is worthwhile to ask how large do sample sizes need to be before biased regularization becomes effective and whether that large sample size makes regularization unnecessary. There is no single answer, because this work has shown that unbiased regularization always decreases error as sample size increases.

By Result 4, the assumption of manifold diffeomorphism is not practically constraining for large sample sizes if $J_{\partial\theta,\eta}$ exists, because Θ_1 need only model parameter values near θ_1. Of course, θ_1 is never really known, so it may be strategic to choose models with completely differentiable parameter manifolds. It is fair to argue that a *true* model may not be estimatable with a differentiable parameter manifold, so we may need to ask if a completely differentiable manifold could approximate such structures for as $n \to \infty$, or if the diffeomorphic assumption could be relaxed. While relevant to the generalization of the theory produced in this work, such questions quickly leave the realm of applied statistics and become geometric in nature.

This work also illustrates how unregularized models may implicitly behave as if they are regularized. A model's parameter space is often a subspace Θ_0 of some undefined larger set Θ_1. For example, a Poisson distribution is generalized by the Negative Binomial. This work shows that if the model is close enough, such generalizations may not matter. In using statistical models as a proxy for reality, simplicity may be essential to feasibility.

6. Conclusion

This work has demonstrated that already popular penalized likelihood regularization implicitly satisfies the assumption of parameter manifold consistency. In turn, this assumption is proven to decrease estimation error for sufficiently large sample sizes when bias is small. Further, this condition implies the existence of a parameter manifold that can be both biased and have less error than a larger, unbiased manifold. So in applications where a sought *true* model may require infinitely many parameters to estimate, it is more pragmatic to simply be close enough.

7. References

[1] Hirotuga Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] Peter Bickel, Bo Li, Alexandre Tsybakov, Sara Geer, Bin Yu, Teófilo Valdés, Carlos Rivero, Jianqing Fan, and Aad Vaart. Regularization in statistics. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 15(2):271–344, 2006.

[3] Tony Cai, Weidong Liu, and Xi Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *American Statistical Association*, 106(494):594–607, 2011.

[4] Roger R. Davidson and William E. Lever. The limiting distribution of the likelihood ratio statistic under a class of local alternatives. *Sankhyā: The Indian Journal of Statistics*, 32(2):209–224, 1970.

[5] W. Evan Durno. *Precise correlation and metagenomic binning uncovers fine microbial community structure*. UBC (Master's thesis). Retrieved from https://circle.ubc.ca/, 2017.

[6] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2008.

[7] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000.

[8] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

[9] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, 231:289–337, 1933.

[10] Mohsen Pourahmadi. *High-Dimensional Covariance Estimation*. Wiley, 2013.

[11] R Core Team. Writing r extensions, 2016. cran.r-project.org/doc/manuals/r-release/R-exts.html;accessed online 4 Oct 2016.

[12] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.