Putting regular statistical models into standard form

Kaleb D. Ruscitti

A central theorem in singular learning theory (SLT) is that every singular analytic statistical model can be put into a standard form. In this form, the learning coefficients of the model can be read off immediately.

In the SLT literature, a common example given is that regular models of dimension d have learning coefficient $d/2$ with multiplicity 1. This is usually stated without further details, but this fact is not entirely obvious. It can be obtained in multiple ways, including by putting the regular model into standard form.

As an exercise in understanding the standard form theorem better, in this article we perform a blow-up to resolve the regular model into standard form. Once in standard form, we can easily obtain the learning coefficient $d/2$ and its multiplicity 1.

1 Singular Standard Form

Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, and let $\mathcal{P}(\Omega)$ be the set of probability density functions on Ω. For our purposes, a (Bayesian) *statistical model* for Ω is a triple (W, p, ϕ), where $W \subset \mathbb{R}^d$ is a measurable set of parameters, $p : W \to \mathcal{P}(\Omega)$, and $\phi \in \mathcal{P}(W)$ is a prior distribution on W.

Singular learning theory is concerned with models that satisfy some analyticity conditions that allow us to apply ideas from algebraic geometry. We'll call a statistical model *analytic* if:

1. W is a semi-analytic subset of $\mathbb{R}^d$, meaning there exists analytic functions $f_1, ..., f_k : \mathbb{R}^d \to \mathbb{R}$ such that
$$W = \{w \in \mathbb{R}^d \mid f_i(w) \geqslant 0 \ \forall i = 1, ..., k\}.$$
2. For any two points $w_1, w_2 \in W$, the Kullback-Leibler divergence
$$K(w_1|w_2) := K(p(x|w_1)|p(x|w_2) \tag{1.1}$$
is analytic in both parameters.
3. The prior can be decomposed as $\phi = \phi_s \phi_a$, where ϕ_s is smooth with $0 < \phi_s < \infty$ and ϕ_a is analytic with $0 \leqslant \phi_a < \infty$.

One major result in SLT is the standard form theorem of Watanabe's, which says that every statistical model can be reparametrized into a standard form. In his textbook *Algebraic Geometry and Statistical Learning Theory*, Watanabe calls this Main Theorem 1 [Wat09].

Two pieces of notation: First, for any subset $I \subset \{1, ..., d\}$, any vector $v = (v_1, ..., v_d) \in \mathbb{R}^d$, and any vector of integers $(k_1, ..., k_d)$, let
$$v_I^{k_I} := \prod_{i \in I} v_i^{k_i}.$$
Second, for any choice of density $q(x) \in \mathcal{P}(\Omega)$, define $K_q(w) : W \to \mathbb{R}$ as
$$K_q(w) = K(p(x|w) \mid q(x)).$$

Definition 1 (Standard Form). Let $I = \{1, ..., d\}$. An analytic statistical model (W, p, ϕ) is in standard form (with respect to q) if there exists a basis $w = (w_1, .., w_d)$ of $\mathbb{R}^d$ and integers $(k_i, h_i)_{i=1}^d$, such that:

$$K_q(w) = w_I^{2k_I} + \mathcal{O}(w_I^{2k_I+1}) \qquad \text{and} \qquad \phi(w) = \phi_s(w) \cdot |w_I^{h_I}|,$$

where $\phi_s(w)$ is smooth and $\phi_s(w) \neq 0$ everywhere.

Watanabe proved that by using Hironaka's Theorem for the resolution of singularities in characteristic zero, any statistical model can be transformed into standard form by applying a birational map. One reason for doing this is to compute the learning coefficient of the model using the following theorem:

Theorem 1. Suppose a model is in standard form with respect to q with integers $(k_i, h_i)_{i=1}^d$. Then the learning coefficient of the model w.r.t. q is

$$\min_{i=1,...,d} \frac{h_i + 1}{2k_i},$$

with multiplicity given by the number of times the minimum is achieved.

Proof. [Mus11, Theorem 1.1] □

It is well-known that regular statistical models have learning coefficients $\dim W/2$ with multiplicity 1. This can be seen without applying the standard form theorem, but it is natural to ask what the process of computing the learning coefficients by resolving a regular model looks like.

For this article, a *regular* statistical model (W, p, ϕ) means:

1. W is a d-dimensional submanifold of $\mathbb{R}^d$,
2. For every $x \in \Omega$, the function $p(x|w)$ is smooth in w.
3. The map $p : W \to \mathcal{P}(\Omega)$ is injective.
4. The Fisher information $I(w)$ is positive-definite everywhere.
5. The prior satisfies $0 < \phi(w) < \infty$.

2 Projective Space

Let us briefly detour from statistics to introduce key facts about projective space that we need to perform a blow-up. The projective space $\mathbb{P}(\mathbb{R}^d)$ is the set of lines through the origin in $\mathbb{R}^d$. It can be expressed in many ways, but here is a simple one: for $x, x' \in \mathbb{R}^d$ define an equivalence relation $x \sim x'$ if there exists $\lambda \neq 0 \in \mathbb{R}$ such that $x' = \lambda x$. Then $\mathbb{P}(\mathbb{R}^d) = (\mathbb{R}^d \setminus \vec{0})/\sim$.

Let $(z_1, ..., z_d)$ be a basis for $\mathbb{R}^d$. Then since $\mathbb{P}(\mathbb{R}^d) = (\mathbb{R}^d \setminus \vec{0})/\sim$, any point $z \in \mathbb{P}(\mathbb{R})^d$ can be represented by a point $(z_1, ...z_d) \in \mathbb{R}^d$. Traditionally one writes

$$z = [z_1 : z_2 : \cdots : z_d]$$

and refers to this as *homogeneous co-ordinates* for z. Be careful: by definition, for any $\lambda \neq 0 \in \mathbb{R}$,

$$[z_1 : z_2 : \cdots : z_d] = [\lambda z_1 : \lambda z_2 : \cdots : \lambda z_d],$$

so the choice of homogeneous co-ordinates for a point is only unique up to scalar multiplication.

Next lets see that projective space $\mathbb{P}(\mathbb{R}^d)$ is a $(d-1)$-dimensional smooth manifold. For each $i = 1, ..., d$, define an open set $U_i \subset \mathbb{P}(\mathbb{R}^d)$ as

$$U_i = \{[z_1 : z_2 : \cdots : z_d] \in \mathbb{P}(\mathbb{R})^d \mid z_i \neq 0\}.$$

Then we have a bijective chart $t^{(i)} : U_i \to \mathbb{R}^{d-1}$ by

$$t^{(i)}([z_1 : \cdots z_d]) = (z_j/z_i)_{j \neq i} = \left(\frac{z_1}{z_i}, \cdots \frac{z_d}{z_i}\right).$$

The $\cdots$ in the last expression skips the ith entry z_i/z_i; i.e. $t^{(i)}_j = z_j/z_i$ for $j \neq i$. The collection $\{(U_i, t^{(i)})\}_{i=1}^d$ defines an atlas of co-ordinate charts that cover $\mathbb{P}(\mathbb{R}^d)$, giving $\mathbb{P}(\mathbb{R}^d)$ a smooth manifold structure. These charts are called the *standard charts* for projective space. If you're learning geometry, it is a good and common exercise to check that the transition functions of these charts are indeed smooth.

In the case $d = 2$, the manifold $\mathbb{R}(\mathbb{P}^2)$ is called the real projective line and it is diffeomorphic to the circle S^1. I want to treat it a bit differently, as this treatment will let us draw a picture of a blow-up in the next section. Let $z = [z_1 : z_2]$. The set U_1, where $z_1 \neq 0$, covers every point in $\mathbb{R}(\mathbb{P}^2)$ except the point $z = [0 : 1]$. The chart $t^{(1)} : U_1 \to \mathbb{R}$ is

$$t^{(1)}([z_1 : z_2]) = z_2/z_1,$$

so at the one point where $z_1 = 0$, we simply define $t^{(1)}([0 : 1]) = \infty$. This defines a bijective map $\mathbb{R}(\mathbb{P}^2) \to \mathbb{R} \cup \{\infty\}$ that lets us identify $\mathbb{R}(\mathbb{P}^2) = \mathbb{R} \cup \{\infty\}$.

There is geometric intuition for this identification. Recall that $\mathbb{R}(\mathbb{P}^2)$ is defined to be the set of lines through the origin in $\mathbb{R}^2$. If we treat the vertical line $x = 0$ as having a slope of ∞, then every line is given by an equation $y = mx$, for some $m \in \mathbb{R} \cup \{\infty\}$. Thus identifying $\mathbb{R}(\mathbb{P}^2) = \mathbb{R}^2 \cup \{\infty\}$ is parametrizing lines in $\mathbb{R}^2$ by their slope.

3 Resolving a Regular Model

Let $M = (W, p, \phi)$ be a regular analytic statistical model for a probability space Ω. Fix a probability density $q \in \mathcal{P}(\Omega)$, and suppose that there is $w_0 \in W$ such that $p(x|w_0) = q$. Since M is regular, this w_0 must be the unique global minimum of $K_q(w)$, satisfying $K_q(w_0) = 0$. It is well-known that the Hessian of a regular model is the Fisher information. As w_0 is the global minimum of $K_q(w)$, $\nabla K_q(w) = 0$. Using these facts, the Taylor expansion of $K_q(w)$ near w_0 is

$$K_q(w) = (w - w_0)^T I(w)(w - w_0) + \mathcal{O}(w^3).$$

By choosing a basis for $\mathbb{R}^d$ that diagonalizes $I(w)$ and translating it by $-w_0$ we can write this as

$$K_q(w) = \sum_{i=1}^{d} \alpha_i w_i^2 + \mathcal{O}(w^3),$$

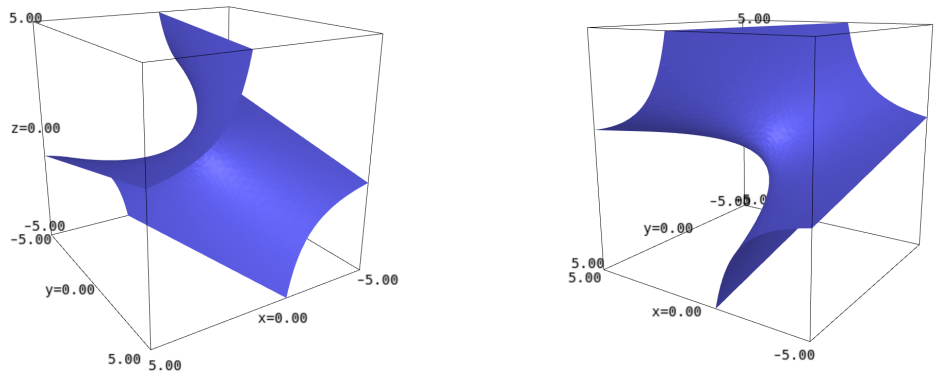
where α_i are the eigenvalues of $I(w)$, which are positive as M is regular.

This model is not in standard form, but we can use a blow-up to reparametrize it into standard form. The blow-up of W at the point w_0 is:

$$\mathrm{Bl}_{w_0} W = \left\{((w_1, ..., w_n, [z_1 : ... : z_n]) \in W \times \mathbb{P}(\mathbb{R}^d) \;\middle|\; w_i z_j = w_j z_i, \forall i, j\right\}.$$

This is a d-dimensional smooth manifold, and the standard charts for $\mathbb{R}(\mathbb{P}^d)$ give us charts for $\mathrm{Bl}_{w_0} W$. Let $V_i \subset \mathrm{Bl}_{w_0} W$ be

$$V_i = \{(w, z) \in \mathrm{Bl}_{w_0} W \mid z_i \neq 0\}, \tag{3.1}$$



(a) One perspective.

(b) Another perspective.

Figure 1: Two perspectives of $\mathrm{Bl}_{w_0}\mathbb{R}^2 \cap [-5,5]^3$, plotted using SageMath. An interactive version of this plot is available at kaleb.ruscitti.ca/blowup.

and let $s^{(i)}(w,z)$ be the vector in $\mathbb{R}^d$ with

$$s^{(i)} = \begin{cases} t_j^{(i)} = \frac{z_j}{z_i}, & j \neq i, \\ w_i, & j = i. \end{cases} \tag{3.2}$$

The collection $\{(V_i, s^{(i)})_{i=1}^d\}$ defines an atlas for $\mathrm{Bl}_{w_0}W$.

To get a better feeling for the blow-up, lets fix $d = 2$ and $W = \mathbb{R}^2$, then take a moment to draw a picture of $\mathrm{Bl}_{w_0}\mathbb{R}^2$. For $d = 2$,

$$\mathrm{Bl}_{w_0}W = \left\{(w_1, w_2, [z_1 : z_2]) \in W \times \mathbb{P}(\mathbb{R}^2) \;\middle|\; w_1 z_2 = w_2 z_1\right\},$$

and there are two charts, $V_1 \cong \mathbb{R}^2$ with co-ordinates $(w_1, z_2/z_1)$ and $V_2 \cong \mathbb{R}^2$ with co-ordinates $(w_2, z_1/z_2)$. Using our identification of $\mathbb{P}(\mathbb{R}^2) = \mathbb{R} \cup \{\infty\}$, we can draw a picture of almost all of $\mathrm{Bl}_{w_0}\mathbb{R}^2$ by dropping only the points $(w,z) \in \mathrm{Bl}_{w_0}\mathbb{R}^2$ with $z = \infty$. We call such points "points at infinity". Now we have

$$\left(\mathrm{Bl}_{w_0}\mathbb{R}^2 - \{\text{points at infinity}\}\right) \subset \mathbb{R}^3,$$

so we can draw this portion of $\mathrm{Bl}_{w_0}\mathbb{R}^2$ in a 3D plot! Specifically, using co-ordinates (w_1, w_2, t), where $t = z_2/z_1$, we have that $w_1 z_2 = w_2 z_1$ becomes $w_1 t = w_2$. Therefore, $\mathrm{Bl}_{w_0}\mathbb{R}^2$ is the surface defined by the equation $w_1 t = w_2$ inside $\mathbb{R}^3$ (Figure 1).

Remember, this picture is missing one small bit of $\mathrm{Bl}_{w_0}\mathbb{R}^2$. The points at infinity are points with $z_1 = 0$, which must satisfy $w_1 z_2 = w_2 z_1 = 0$. There is an $\mathbb{R}$-worth of these points, given by the choice of w_2. So our picture is missing a *line at infinity*. The picture is called one *affine slice* of the blow-up. There is another affine slice corresponding to the other chart V_2 for $\mathrm{Bl}_{w_0}W$. In general dimension d, there will be d-many standard affine slices of $\mathrm{Bl}_{w_0}W$, and each slice misses a $(d-1)$-dimensional hyperplane at infinity.

The blow-up comes with an analytic map:

$$\pi : \mathrm{Bl}_{w_0}W \to W, \qquad (w,z) \mapsto w.$$

Using this map, we can define a model $\tilde{M}$ as a reparametrization of M. First, let $\pi^* p : \mathrm{Bl}_{w_0}W \to \mathcal{P}(\Omega)$ be

$$(\pi^* p)(x|w,z) := p(x|\pi(w,z)) = p(x|w).$$

Second, let $\pi^*\phi \in \mathcal{P}(\mathrm{Bl}_{w_0}W)$ be

$$(\pi^*\phi)(w,z) = \phi(\pi(w,z)) \cdot |\det J_\pi(w,z)| = \phi(w) \cdot |\det J_\pi(w,z)|,$$

where J_π is the Jacobian matrix of π. Then $\tilde{M} = (\mathrm{Bl}_{w_0} W, \pi^* p, \pi^* \phi)$ is an analytic statistical model, and our goal is to show that it is in standard form.

For the original model M we saw that the Kullback-Leibler divergence is approximately

$$K_q(w) = \sum_{i=1}^{d} \alpha_i w_i^2,$$

where $\alpha_i > 0$. Our new model $\tilde{M}$ has Kullback-Leibler $\pi^* K_q(w, z) = K_q(\pi(w, z))$, however we want to write this in our co-ordinate charts for $\mathrm{Bl}_{w_0} W$. On the open set V_i with co-ordinates $s^{(i)} = (w_i, t^{(i)})$ we have

$$\pi^* K_q(w_i, t^{(i)})|_{V_i} = \alpha_i w_i^2 + \sum_{j \neq i} \alpha_j (t_j^{(i)} w_i)^2 = w_i^2 \left(\alpha_i + \sum_{j \neq i} \alpha_j (t_j^{(i)})^2 \right).$$

To see that this is in standard form, we replace w_i with u_i,

$$u_i := w_i \sqrt{\alpha_i + \sum_{j \neq 1} \alpha_j (t_j^{(i)})^2}.$$

This is always well-defined because $\alpha_j > 0$ for every $j = 1, ..., d$. In $(u_i, t^{(i)})$ co-ordinates, we have

$$\pi^* K_q(u_i, t^{(i)})|_{V_i} = u_i^2.$$

Thus, the Kullback-Leibler divergence of $\tilde{M}$ is in standard form.

Next, we need to check that the prior is in standard form. By assumption $0 < \phi(w) < \infty$ everywhere on W, and by definition

$$\pi^* \phi(w, z) = \phi(\pi(w, z)) \cdot |\det J_\pi(w, z)|.$$

So we need to compute $J_\pi(w, z)$. In the co-ordinate chart V_i, we have

$$\begin{aligned} J_\pi(w_i, t^{(i)})|_{V_i} &= \begin{bmatrix} \frac{\partial w_i}{\partial w_i} & \frac{\partial}{\partial w_i}(w_i t_1^{(i)}) & \cdots & \frac{\partial}{\partial w_i}(w_i t_d^{(i)}) \\ \frac{\partial w_i}{\partial t_1^{(i)}} & \frac{\partial}{\partial t_1^{(i)}}(w_i t_1^{(i)}) & \cdots & \frac{\partial}{\partial t_i}(w_i t_d^{(i)}) \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial w_i}{\partial t_d^{(i)}} & \frac{\partial}{\partial t_d^{(i)}}(w_i t_1^{(i)}) & \cdots & \frac{\partial}{\partial t_d^{(i)}}(w_i t_d^{(i)}) \end{bmatrix} \\ &= \begin{bmatrix} 1 & t_1^{(i)} & \cdots & t_d^{(i)} \\ 0 & w_i & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & w_i \end{bmatrix}. \end{aligned}$$

This is upper triangular, so its determinant is the product of its diagonal entries: $\det J_\pi(w_i, t_i)|_{V_i} = w_i^{d-1}$. Therefore the prior on $\mathrm{Bl}_{w_0} W$ is

$$\pi^* \phi(w_i, t^{(i)})|_{V_i} = \phi(w) \cdot |w_i^{d-1}|.$$

However for a model to be in standard form, we need to express the Kullback-Leibler divergence and the prior in the same co-ordinates. Therefore, we want to re-write the prior using the co-ordinate u_i from our discussion of the Kullback-Leibler divergence. Let $(\alpha_i + \sum_{j \neq 1} \alpha_j (t_j^{(i)})^2)^{-1/2} =: a(t^{(i)})$, so that $u_i = w_i / a(t^{(i)})$. Since the eigenvalues α_j are positive, $a(t^{(i)}) \neq 0$, so this is well defined. In co-ordinates $(u_i, t^{(i)})$ for V_i, we have

$$\begin{aligned} \pi^* \phi(u_i, t^{(i)}) &= \phi(\pi(u_i, t^{(i)})) \cdot a(t^{(i)})^{d-1} \cdot |u_i^{d-1}|, \\ \pi^* K_q(u_i, t^{(i)}) &= u_i^2. \end{aligned}$$

Thus we can see that the model $\tilde{M}$ is in standard form. Moreover, we can read off that the integers (k_j, h_j) which determine the learning coefficient are

$$k_j = \begin{cases} 1, & j = i, \\ 0, & \text{else} \end{cases} \qquad h_j = \begin{cases} d-1, & j = i, \\ 0, & \text{else.} \end{cases} \tag{3.3}$$

Therefore, using Theorem 1 we obtain that the learning coefficient is $\lambda = d/2$ and its multiplicity is $m = 1$.

References

[Mus11] Mircea Mustata. *IMPANGA lecture notes on log canonical thresholds*. July 13, 2011. DOI: 10.48550/arXiv.1107.2676.

[Wat09] Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009. ISBN: 987-0-521-86467-1.