Bernoulli **19**(4), 2013, 1449–1464
DOI: 10.3150/12-BEJSP15

The geometry of least squares in the 21st century

JONATHAN TAYLOR

Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, CA 94305, USA. E-mail: jonathan.taylor@stanford.edu

It has been over 200 years since Gauss's and Legendre's famous priority dispute on who discovered the method of least squares. Nevertheless, we argue that the normal equations are still relevant in many facets of modern statistics, particularly in the domain of high-dimensional inference. Even today, we are still learning new things about the law of large numbers, first described in Bernoulli's *Ars Conjectandi* 300 years ago, as it applies to high dimensional inference.

The other insight the normal equations provide is the asymptotic Gaussianity of the least squares estimators. The general form of the Gaussian distribution, Gaussian processes, are another tool used in modern high-dimensional inference. The Gaussian distribution also arises via the central limit theorem in describing weak convergence of the usual least squares estimators. In terms of high-dimensional inference, we are still missing the right notion of weak convergence.

In this mostly expository work, we try to describe how both the normal equations and the theory of Gaussian processes, what we refer to as the "geometry of least squares," apply to many questions of current interest.

Keywords: convex analysis; Gaussian processes; least squares; penalized regression

1. Basic tools in the geometry of least squares

The method of least squares has by now a long and well-trodden history, which we will not attempt to address in this work. Toward the end of the 20th century, Stigler (1981) referred to the method of least squares as the automobile of (then) modern statistical analysis. Today, 30 years later, as automobiles have modernized, including technological and efficiency improvements, so too have the methods of least squares changed.

Let us recall the classical least squares problem: given outcome vector $y \in \mathbb{R}^n$ and design matrix $X \in \mathbb{R}^{n \times p}$, the least squares problem is typically posed as

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \frac{1}{2} \|y - X\beta\|_2^2. \tag{1}$$

The related normal equations are

$$X^T (X\hat{\beta} - y) = 0. \tag{2}$$

We could have almost equivalently written the equations above in terms of the metric projection problem

$$y \mapsto \operatorname*{argmin}_{x \in A} \frac{1}{2}\|y - x\|_2^2 \stackrel{\Delta}{=} \pi_A(y) \tag{3}$$

with classical least squares by setting $A = \mathrm{col}(X)$ the column space of X. This metric projection problem is the first of the basic tools in what we will refer to here as "the geometry of least squares in the 21st century."

While written as a function above, the map (3) is not always a function, there may be many minimizers for a given y. For a set A, define the *critical radius* of A as

$$r_c(A) = \sup\Big\{r : \inf_{x \in A} \|y - x\|_2 \leq r \Longrightarrow (3) \text{ has a unique solution}\Big\}. \tag{4}$$

Note that for convex A, $r_c(A) = +\infty$. The metric projection problem also makes sense for other sets and other metrics. For instance, suppose $A \subset S(\mathbb{R}^n) = S_{\ell_2}(\mathbb{R}^n)$, the unit ℓ_2 sphere in $\mathbb{R}^n$ with distance $d(x, y) = \cos^{-1}(x^T y)$. One might also consider the spherical metric projection

$$S\big(\mathbb{R}^n\big) \ni y \mapsto \operatorname*{argmax}_{x \in A} x^T y \tag{5}$$

with the critical radius (4) being similarly defined. We will see in Section 4 that the above critical radius plays a part in the supremum of a class of Gaussian processes [Adler and Taylor (2007)], one of the other important class of objects associated to Gauss's name. Gaussian processes suffer some of the same deficiencies identified in Stigler (1981): they make many assumptions and have their limitations. Nevertheless, they are a crucial inferential tool in analyzing the behavior of (3). Hence, we refer to these as the second of the basic tools in the geometry of least squares in the 21st century.

2. A canonical high-dimensional regression problem

In the classical setting $n > p$ the system (2) often has a unique solution, the familiar

$$\hat{\beta} = \big(X^T X\big)^{-1}\big(X^T y\big).$$

In many parametric models, the least squares model is of course too simple. In the exponential family setting [Amari and Nagaoka (2000), Efron (1978)], the normal equations are similar, with $(X^T X)$ replaced by the observed Fisher information. We have focused on squared error-loss for its simplicity of exposition.

In high-dimensional settings, n is often less than p and there is of course no unique solution to (2). Many modifications are possible, for instance, ridge or Tikhonov regularization which adds a strongly convex quadratic term to (1). The addition of such quadratic terms changes the quadratic part of the loss but does not fundamentally change much else until we begin to make assumptions about whether or not the model is correct, and how much bias might be incurred by such regularization.

In modern high-dimensional settings the regularization term, or penalty, of choice is often a norm, with the LASSO [Tibshirani (1996)] being the most popular. The lasso problem is

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1. \tag{6}$$

The duality between norms allows us to write

$$\|\beta\|_1 = \sup_{u:\|u\|_\infty\le 1} u^T\beta = h_{B_\infty}(\beta)$$

with B_∞ the ℓ_∞ ball of radius 1 in $\mathbb{R}^p$ and for any set K

$$h_K(\beta) = \sup_{u\in K} u^T\beta \tag{7}$$

is the support function of the set K which we assume to be closed and containing 0. In this notation, the LASSO problem can be expressed as

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda h_{B_\infty}(\beta). \tag{8}$$

Our canonical problem is therefore

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda h_K(\beta) \tag{9}$$

with K being a closed, convex set containing 0. For one of many possible infinite dimensional formulations of this canonical problem, see Tsirel'son (1982) whose author is also associated to one of the most famous tools in the theory of Gaussian process, the Borell TIS inequality [Adler and Taylor (2007)].

The normal equations of the least squares problem are replaced with the KKT conditions [Boyd and Vandenberghe (2004)] for (9). For our canonical problem, the KKT conditions are

$$X^T(X\hat\beta - y) + \hat u = 0, \qquad \hat u \in \lambda\cdot\partial h_K(\hat\beta), \tag{10}$$

where the ∂ denotes the sub differential. In what follows, we denote a solution to this problem as $\hat\beta_{\lambda,K_X}$ to denote the dependence on the penalty K_X and the penalty parameter λ.

As we can encode linear or cone constraints in the support function, it is safe to say that a huge number of problems fit into this framework. Some examples include:

- LASSO [Tibshirani (1996)];
- compressed sensing [Candès, Romberg and Tao (2006), Donoho (2006)];
- group LASSO [Yuan and Lin (2006), Obozinski, Jacob and Vert (2011)];
- graphical LASSO [Friedman, Hastie and Tibshirani (2008), Bühlmann (2012)];
- matrix completion [Candès and Recht (2009), Mazumder, Hastie and Tibshirani (2010)];
- sign restricted regression [Meinshausen (2012)];

- hierarchically constrained models [Bien, Taylor and Tibshirani (2013), Jenatton *et al.* (2011)];
- generalized LASSO [Tibshirani and Taylor (2011)];
- total variation denoising [Becker, Bobin and Candès (2011)].

The relation between (9) and the metric projection map is through a particular dual function, also developed by Legendre. The canonical dual problem is

$$\underset{u\in\lambda\cdot\mathrm{row}(X)\cap K}{\text{minimize}}\ \frac{1}{2}(u-w)^T\left(X^TX\right)^{\dagger}(u-w) \tag{11}$$

with $\mathrm{row}(X)$ the row space of the matrix X. This dual problem can be derived by minimizing the following Lagrangian with respect to β, η

$$L(\beta,\eta;u)=\tfrac{1}{2}\|y-X\beta\|_2^2+\lambda h_K(\eta)+u^T(\beta-\eta). \tag{12}$$

After a sign change, the problem in (11) is fairly easily seen to be equivalent to

$$\underset{u}{\text{maximize}}\left[\inf_{\beta,\eta}L(\beta,\eta;u)\right] \tag{13}$$

with the constraints in (11) encoding the fact that

$$\inf_{\beta,\eta}L(\beta,\eta;u)=-\infty$$

whenever $u\notin\lambda\cdot\mathrm{row}(X)\cap K$. Any pair $\hat\beta,\hat u$ is related through

$$\left(X^TX\right)\hat\beta-w+\hat u=0. \tag{14}$$

Choosing an orthonormal basis for $\mathrm{row}(X)$, we see that the dual problem can be phrased as the metric projection problem of $(X^TX)^{-\dagger/2}w$ onto $\lambda\cdot\mathrm{row}(X)\cap K$. Alternatively, if we are interested only in the fitted values $X\hat\beta_{\lambda,K}$, the original problem (9) can be expressed as the residual $\hat\mu=y-\hat r$ from

$$\hat r=\underset{r\in\lambda\cdot K_X}{\mathrm{argmin}}\ \frac{1}{2}\|y-r\|_2^2, \tag{15}$$

where

$$K_X=\left(X^T\right)^{-1}K. \tag{16}$$

Or, in another form,

$$\hat\mu(y)=y-\pi_{\lambda\cdot K_X}(y). \tag{17}$$

We see that our canonical regression problem is in fact a metric projection problem.

Having posed the canonical high-dimensional regression problem as a metric projection problem, we now try to describe how metric projection is related to some fundamental issues in the understanding of this problem both from an algorithmic view and an inferential view.

2.1. Algorithms to solve the canonical problem

The general problem is phrased in terms of an arbitrary K. For many high-dimensional problems, this K is chosen to emphasize expected structure in the data. For example, it is well known that the LASSO yields sparse solutions, the group LASSO yields groups of nonzero coefficients, etc.

This special structure is based on a particular structure encoded in K. Further, for many canonical choices of K used in high dimensional statistics, such as those cited in Section 2, the following metric projection is simple

$$\nu \mapsto \operatorname*{argmin}_{u \in \lambda \cdot K} \frac{1}{2}\|\nu - u\|_2^2 = \nu - \operatorname*{argmin}_{\beta} \left[\frac{1}{2}\|\nu - \beta\|_2^2 + \lambda h_K(\beta) \right].$$

Such optimization problems are referred to as problems in composite form [Becker, Bobin and Candès (2011), Boyd and Vandenberghe (2004), Nesterov (2005)]. For such problems, many modern first order solvers use a version of generalized gradient descent. The steps in generalized gradient descent are essentially iterations of this metric projection map. Specifically, to solve the canonical problem (9) for a given step-size α a simple generalized gradient algorithm reads as

$$\begin{aligned} \hat{\beta}^{(k+1)} &= \operatorname*{argmin}_{\beta} \frac{1}{2\alpha} \left\| \nu_\alpha^{(k)} - \beta \right\|_2^2 + \lambda h_K(\beta) \\ &= \nu_\alpha^{(k)} - \operatorname*{argmin}_{\eta \in \lambda\alpha \cdot K} \frac{1}{2} \left\| \nu_\alpha^{(k)} - \eta \right\|_2^2, \\ \nu_\alpha^{(k)} &= \beta^{(k)} - \alpha \cdot X^T \left(X\beta^{(k)} - y \right). \end{aligned} \tag{18}$$

The first line in the update above is the usual form of updates for generalized gradient descent, while the second line expresses this step as the residual after an application of the metric projection map. Accelerated schemes can do much better with slightly different updates above, see Becker, Bobin and Candès (2011), Nesterov (2005), Tseng (2013). With modern computing techniques, such simple algorithms can scale to huge problems, see Boyd, Parikh and Chu (2011), Mazumder, Hastie and Tibshirani (2010).

3. Inference for the canonical problem

3.1. Law of large numbers

Having solved the canonical problem, what can we say about its solution? As this special issue is devoted to the appearance of one of the first proofs of the law of large numbers, we should at least hope to provide such an answer.

In the classical setting, assuming independence, and the usual linear regression model

$$y = \mu + \varepsilon \tag{19}$$

with noise ε having scale σ, the central limit theorem can often be applied to (1) yielding the usual result

$$\left\|X(\hat{\beta}-\beta_0)\right\|_2^2 \overset{D}{\approx} \sigma^2 \cdot \chi_p^2 \tag{20}$$

under the null $H_0 : \mu = X\beta_0 \in \mathrm{col}(X)$. Of course, this forms the basis of much inference in modern (and not so modern) applied statistics in the fixed p, n growing regime. In terms of the parameters themselves, this implies the weaker statement

$$\|\hat{\beta}-\beta\|_2 \leq n^{-1/2}\sigma \cdot \zeta_n \tag{21}$$

for some random variable $\zeta_n = O_{\mathbb{P}}(1)$.

In the classical setting, assuming X is full rank, the bound (21) is a simple two line proof followed by some assertions. If we write $\mathcal{L}(\beta) = \frac{1}{2}\|y - X\beta\|_2^2$, then

$$\begin{aligned}
0 &\geq \mathcal{L}(\hat{\beta}) - \mathcal{L}(\beta_0) \\
&= \nabla\mathcal{L}(\beta_0)^T(\hat{\beta}-\beta_0) + \frac{1}{2}(\hat{\beta}-\beta_0)^T\left(X^TX\right)(\hat{\beta}-\beta_0) \\
&= \left(X^T\varepsilon\right)^T(\hat{\beta}-\beta_0) + \frac{1}{2}(\hat{\beta}-\beta_0)^T\left(X^TX\right)(\hat{\beta}-\beta_0) \\
&\geq -\left\|X^T\varepsilon\right\|_2\|\hat{\beta}-\beta_0\|_2 + \frac{\lambda_{\min}(X^TX)}{2}\|\hat{\beta}-\beta_0\|_2^2
\end{aligned}$$

with $\lambda_{\min}(X^TX)$ denoting the smallest eigenvalue of X^TX. We see that we can take

$$\zeta_n = \sigma^{-1/2}n^{1/2}\frac{\|X^T\varepsilon\|_2}{\lambda_{\min}(X^TX)}.$$

In the high-dimensional setting, of course this fails as $\lambda_{\min}(X^TX) = 0$. What, then, can we say about our canonical estimator

$$\hat{\beta}_{\lambda,K_X} = \operatorname*{argmin}_{\beta\in\mathbb{R}^p}\frac{1}{2}\|y - X\beta\|_2^2 + \lambda h_K(\beta)?$$

Is there even a weak law of large numbers? Without any assumptions on K, the answer is clearly no: if $K = \{0\}$, then this is the original ill-posed problem.

Under an assumption of decomposability of K recent progress has been made in providing bounds on the estimation error in (9), see Negahban *et al.* (2012). The notion of decomposability in Negahban *et al.* (2012) has a precise definition which we will not dwell on here. However, a large set of examples of decomposable penalties are penalties of the form

$$K = \prod_{i\in\mathcal{I}} K_i. \tag{22}$$

That is, convex sets that can be expressed as products of convex sets lead to decomposable penalties. Every example of a decomposable norm in Negahban *et al.* (2012) has this form except

the nuclear norm. For such penalties, the generalized gradient algorithms described in Section 2.1 decompose into many smaller subproblems. Many efficient coordinate descent algorithms exploit this fact, see Friedman *et al.* (2007), Friedman, Hastie and Tibshirani (2009).

A similar notion of decomposability we refer to as *additivity* is explored in Taylor and Tibshirani (2013) in which case K can be expressed as $K = A + I$ with the sum being Minkowski addition of sets. In this case, the penalty has the form

$$h_K(\beta) = h_A(\beta) + h_I(\beta). \tag{23}$$

One concrete example of this is the ℓ_∞ ball in $\mathbb{R}^p$. For A, I a partition of $\{1, \ldots, p\}$ into *active* and *inactive* variables, we can write

$$B_\infty = \big\{(u_A, 0) : \|u_A\|_\infty \leq 1\big\} + \big\{(0, u_I) : \|u_I\|_\infty \leq 1\big\}.$$

Any penalty of the form (22) can be expressed in the form (23) in a similar fashion. If we are allowed to introduce a linear constraint to (9), then any problem with a penalty of the form (23) can be expressed as a problem with a penalty of the form (22) subject to an additional set of linear constraints.

The weak law of large numbers presented in the literature follow a similar path to the argument above. Of course, for precise results, the specific penalty as well as the data generating mechanism must be more precisely specified.

In the interest of space, we do not pursue such precise statements here. Rather, we will just attempt to paraphrase these results, of which there exist many in the literature [cf. Negahban *et al.* (2012), Bickel, Ritov and Tsybakov (2009), Obozinski, Wainwright and Jordan (2011), Bühlmann and van de Geer (2011)] with Negahban *et al.* (2012) being a particularly nice place to read in detail. Under various assumptions on the tails of ε, as well as the assumption $h_I(\beta_0) = 0$ and K° is bounded,[1] the canonical result has the form for $\lambda \geq C_1 \cdot \mathbb{E}(h_{K^\circ}(-X^T\varepsilon))$

$$\|\hat{\beta}_{\lambda, K_X} - \beta_0\|_2 \leq C_2 \cdot \sigma \frac{\mathbb{E}(h_{K^\circ}(-X^T\varepsilon))\psi(A)}{\kappa(A, I, X)} \tag{25}$$

with high probability for some universal C_1, C_2 where $\psi(A)$ is referred to as a *compatibility* constant relating the ℓ_2 norm and the h_A seminorm; the quantity $\kappa(A, X)$ replaces $\lambda_{\min}(X^TX)$ and is referred to as a *restricted strong convexity* (RSC) constant.

The literature varies in their assumptions on the noise and the design matrix X. For instance, in the fixed design case one might consider the above probability only with respect to noise, while in a random design setting the dependence of the constants on X are typically expressed with respect to the law that generates the design matrix X.

Having established a bound such as (25), if one considers problems indexed by n, then one can obtain a law of large numbers for the problem (9) so long as the parameters are chosen so

[1] Recall the definition of the polar body of K

$$K^\circ = \big\{v \in \mathbb{R}^p : u^T v \leq 1, \forall u \in K\big\}. \tag{24}$$

For any K, the seminorm h_{K° is the dual seminorm of h_K.

the right-hand side decays to 0 and the bound holds with sufficiently high probability. Again in the interest of space, we refer readers to the literature for more precise statements for specific versions of the problem such as the LASSO. We have tried to give a summary of some of the results related to the canonical problem, though we have barely exposed the tip of the iceberg. Under more specific assumptions much more can be said. For example, see Donoho, Maleki and Montanari (2009).

3.2. Gaussian width and metric projection: Intrinsic volumes

For $K_X^\circ = X^T K^\circ \subset \mathbb{R}^n$, the error (25) depends on the quantity

$$\mathbb{E}\big(h_{K^\circ}\big(-X^T\varepsilon\big)\big) = \mathbb{E}\big(h_{K_X}(-\varepsilon)\big). \tag{26}$$

For fixed X, this quantity is referred to as the Gaussian width of K_X and it also intimately related to our first tool in the toolbox, the metric projection. Specifically, consider the tube of radius r around K_X°. That is,

$$\mathrm{Tube}\big(K_X^\circ, r\big) = \big\{z \in \mathbb{R}^n : \big\|z - \pi_{K_X^\circ}(z)\big\|_2 \le r\big\}. \tag{27}$$

Then, a classical result of Steiner in the case of convex bodies and Weyl in the case of manifolds says that the Lebesgue measure of the tube, assuming K_X° is bounded, can be expressed as

$$\big|\mathrm{Tube}\big(K_X^\circ, r\big)\big|_{\mathbb{R}^n} = \sum_{j=0}^{n} r^j \omega_j \mathcal{L}_{n-j}\big(K_X^\circ\big), \qquad r \le r_c\big(K_X^\circ\big) = \infty, \tag{28}$$

where the $\mathcal{L}_l(K_X^\circ)$ are referred to as the intrinsic volumes of K_X° and $\omega_l = |B_2(1)|_{\mathbb{R}^l}$ is the Lebesgue measure of the unit ℓ_2 ball in $\mathbb{R}^l$. See Adler and Taylor (2007), Federer (1959), Schneider (1993), Weyl (1939) for more details on such volume of tubes formulae. When K_X° is unbounded, Federer's curvature measures [Federer (1959)] can be used to define the volume of local tubular neighborhoods. Using Gaussian process techniques [Vitale (2001)], it can be shown that

$$\mathcal{L}_1\big(K_X^\circ\big) = (2\pi)^{-1/2}\mathbb{E}\big(h_{K_X^\circ}(\varepsilon)|X\big), \tag{29}$$

where $\varepsilon|X \sim N(0, I_{n\times n})$.

For $D = K_X$ a smooth domain, that is convex set with non-empty interior bounded by a smooth hypersurface and for $j \le n-1$

$$\mathcal{L}_j(D) \propto \int_{\partial D} P_{p-j-1}(\lambda_{1,x}, \ldots, \lambda_{p-1,x})\,\mathrm{Vol}_{\partial D}(dx) \tag{30}$$

with P_j the jth elementary symmetric polynomial of the so-called principal curvatures of ∂D at x [Adler and Taylor (2007), Taylor and Worsley (2007)]. These are just the eigenvalues of the second fundamental form in the unit inward normal direction. This formula can be derived by considering the inverse of the metric projection map (3). The inverse takes (x, η_x) defined on the

extended outward normal bundle of K°. The inverse of the map is simply the exponential map restricted to the outward normal bundle, or, more simply

$$x \mapsto x + \eta_x. \tag{31}$$

A fairly straightforward calculation yields the relation (30).

The main take away message above is that the functionals in the tube formula, such as the Gaussian width (26), are related to properties of the metric projection map onto K_X.

As written above, the intrinsic volumes are defined implicitly through a volume calculation, and it is not clear that they extend to the infinite dimensional setting. Under the right conditions of course, such extensions are indeed possible. See Vitale (2001) for a nice discussion of this problem. An alternative definition of intrinsic volumes specific to the Gaussian case was considered in Taylor and Vadlamani (2013), which were defined as coefficients in an expansion of the *Gaussian* measure of $\text{Tube}(K_X^\circ, r)$ as described in the Gaussian Kinematic Formula [Adler and Taylor (2007), Taylor (2006)].

3.3. Risk estimation

Another quantity of interest for our problem (9) is an estimate of how much "fitting" we are performing, as a function of λ. One quantitative measure of this is captured by Stein's estimate of risk [Stein (1981)], also known as SURE. Suppose now that $y \sim N(\mu, I)$ and we estimate μ by the estimator $\hat{\mu}$. The SURE estimate is an unbiased estimate of

$$\text{Risk}(\hat{\mu}) = \mathbb{E}\big(\|\mu - \hat{\mu}\|_2^2\big).$$

The estimated degrees of freedom of this estimator is one part of the SURE estimate and is defined as

$$\widehat{\text{Cov}(y, \hat{\mu})} = \text{div}\big(\nabla \hat{\mu}(y)\big). \tag{32}$$

Suppose $y \sim N(\mu, I)$ and consider our residual form of the original estimation problem for $\mu = \mathbb{E}(y)$

$$\hat{\mu}(y) = y - \pi_{\lambda K_X}(y). \tag{33}$$

If K possesses a nice stratification, as do all the examples mentioned above, then, for almost every y, $\pi_{\lambda K_X}(y)$ is in the relative interior of some fixed stratum $\mathcal{S}$ of the normal bundle of K_X over which the dimension of the tangent space is constant, and the normal bundle has a locally conic structure $\mathcal{S} = \mathcal{T} \times \mathcal{N}$ of tangent and normal directions [Schneider (1993), Adler and Taylor (2007)]. Having fixed this stratum, we can write $y = x + \eta_x$ in (ortho)normal coordinates centered at $(y - \hat{\mu}(y), \hat{\mu}(y))$. In these coordinates

$$\hat{\mu}\big(y(x, \eta_x)\big) = \eta_x. \tag{34}$$

In order to relate the above to the problem (9), one should invert the above chart to find (x, η_x) in terms of y. In the residual form (17) it is easy to show that

$$\begin{aligned}\big(x(y), \eta_x(y)\big) &= \left(\operatorname*{argmin}_{r \in \lambda \cdot K_X} \frac{1}{2}\|y - r\|_2^2, y - \operatorname*{argmin}_{r \in \lambda \cdot K_X} \frac{1}{2}\|y - r\|_2^2\right) \\ &= \big(y - \hat{\mu}(y), \hat{\mu}(y)\big).\end{aligned}$$

The derivatives along the normal directions yield a purely dimensional term, while directions in the tangent directions yield curvature terms. This observation is enough to derive the following form of the degrees of freedom

$$\operatorname{div}\big(\nabla \hat{\mu}(y)\big) = n - \dim(\mathcal{T}_y) + \operatorname{Tr}(S_{(y-\hat{\mu}(y), \hat{\mu}(y))}). \tag{35}$$

Above, $\mathcal{T}_y$ is the tangential part of the stratum containing $x(y)$ and $S_{(x,\mu_x)}$ is the second fundamental form of $\mathcal{T}_y$ in $\lambda \cdot K_X$ as described [Adler and Taylor (2007)]. When K is a polyhedral set, the second term disappears and the degrees of freedom can be computed by computing the rank of a certain matrix [Tibshirani and Taylor (2012), Bien, Taylor and Tibshirani (2013)].

3.4. Hypothesis testing: Weak convergence for high-dimensional inference?

Another fundamental tool in inference for least squares models is the ability to form hypothesis tests, as well as confidence intervals for the "true" mean. Such concepts clearly need a model, which we might take to be the usual model

$$y \sim N(\mu, I).$$

Under the assumption that $\mu = X\beta_0$, classical inference in linear models (assuming $X^T X$ is full rank) yields confidence intervals and hypothesis tests for any linear functional $v^T \mu$ based on the coefficients β_0.

What can we say about our canonical problem (9)? This is an area in which we still don't know all the answers. In some sense, we are in the situation analogous to Bernoulli having proved a weak law of large numbers without the central limit theorem. Some progress has been made for specific models of the design matrix X for the LASSO as well as group LASSO models, see Meinshausen and Bühlmann (2010), Bühlmann (2012), Minnier, Tian and Cai (2011), Laber and Murphy (2011), Wasserman and Roeder (2009), Donoho, Maleki and Montanari (2009).

Other recent work [Lockhart *et al.* (2013)] gives some hints at what inferential tools may prove useful in this weak convergence theory. As described in the LARS algorithm [Efron *et al.* (2004), Tibshirani (2012)], an entire path of solutions $\hat{\beta}_{\lambda, K_X}$ can be formed for the LASSO, that is, when $K = B_\infty$. These paths are piecewise linear, with knots at points where the *active set* changes. The covariance statistic measures some change in the correlation of the fitted values between two knots λ_k and λ_{k+1} in the LASSO path. It has the form

$$T_k = C_k \lambda_k (\lambda_k - \lambda_{k+1}) \tag{36}$$

for some random scaling C_k related to the active set and the variable added to the active set at λ_{k+1}. The form for $k = 1$ is particularly simple: suppose that $\hat{j}_1$ and X is such the first variable in the LARS path, that is,

$$\hat{j}_1 \in \operatorname*{argmax}_j \left| X_j^T y \right|.$$

Then,

$$T_1 = \frac{1}{\sigma^2} y^T X \hat{\beta}_{\lambda_2} = y^T X_{\hat{j}_1} \hat{\beta}_{\lambda_2, \hat{j}_1}. \tag{37}$$

For $k \geq 1$, the form of the test statistic is slightly more complicated, though it can be expressed in terms of $A = A_k$, the active set at step k as well as s_{A_k}, the signs of the active variables at step k as well as the active set A_{k+1} and $s_{A_{k+1}}$, see Section 2.3 of Lockhart *et al.* (2013) for the full expression. For a wide range of (sequences) covariance matrices, if the active set at λ_k already contains all the strong active variables, then it is shown in Lockhart *et al.* (2013) that

$$T_k \xrightarrow{D} \text{Exp}(1). \tag{38}$$

In particular, under the global null $y \sim N(0, \sigma^2 I)$ as long as the design matrix satisfies some minimum growth condition, $T_1 \xrightarrow{D} \text{Exp}(1)$. The main tools used in the above proof relate to the maxima of (discrete) Gaussian processes and the generalization of an argument previously applied to smooth Gaussian processes in Taylor, Takemura and Adler (2005) and Adler and Taylor (2007). Ongoing work suggests that such a limiting distribution will hold for many (sequences) of K and design matrices X.

As for confidence intervals for the parameters related to k strong variables, the relation to extreme values of Gaussian processes suggest that the bias-corrected or relaxed LASSO estimate of the active coefficients will have accurate coverage. This is ongoing work.

4. Smooth Gaussian processes: Relaxing convexity

In the special case that K is a cone, we saw that the distribution of a particular likelihood ratio test could be expressed in terms of the supremum of a Gaussian process indexed by a subset of the sphere. Equivalently, this supremum could be expressed in terms of the metric projection onto the cone K_X. It is well-known that the distribution of this likelihood ratio test statistic is a mixture of χ^2's of varying degrees of freedom. This distribution is sometimes referred to as a $\bar{\chi}_2$ distribution.

The mixture weights can be expressed in terms of the geometry of the set $M = S(X, K) = K_X \cap S(\mathbb{R}^n)$. In particular, it is known [Adler and Taylor (2007), Takemura and Kuriki (1997, 2002)] that if $\varepsilon \sim N(0, I)$ for $u > 0$

$$\mathbb{P}\Big(\sup_{v \in M} \big(v^T \varepsilon\big)^+ > u \Big) = \sum_{j=0}^{\infty} \mathcal{L}_j(M) \rho_j(u), \tag{39}$$

where

$$\rho_j(u) = \begin{cases} \displaystyle\int_u^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt, & j = 0, \\ (2\pi)^{-(j+1)/2} H_{j-1}(u) e^{-u^2/2}, & j \geq 1, \end{cases} \tag{40}$$

and

$$H_j(u) = (-1)^j \frac{\partial^j}{\partial u^j} e^{-u^2/2}$$

are the standard Hermite polynomials. The functions (40) are known as the EC or Euler characteristic densities for a Gaussian field, see Adler and Taylor (2007), Takemura and Kuriki (2002), Worsley (1995), Worsley *et al.* (1996), Taylor and Worsley (2008). While the sum above is written as an infinite sum it terminates at $\dim(M)$.

Note that this implies

$$\mathbb{E}\Big(\sup_{v\in M}\big(v^T\varepsilon\big)^+\Big) = \frac{1}{\sqrt{8\pi}}\mathcal{L}_1(M)$$

which is an analogous way to derive Gaussian width (29).

One of the derivations of the above formula, the so-called *volume of tubes* approach [Siegmund and Zhang (1993), Sun (1993), Takemura and Kuriki (1997)] involves studying the Jacobian of the inverse of the spherical metric projection map (5), that is, the exponential map on $S(\mathbb{R}^n)$ which sends a pair (x, η_x) to

$$\cos\big(\|\eta_x\|\big)\cdot x + \sin\big(\|\eta_x\|\big)\cdot \eta_x.$$

Another approach, the expected Euler characteristic approach [Adler and Taylor (2007), Worsley (1995)] involves counting critical points of the Gaussian process above the level u according to saddle type and applying Morse theory and the Rice–Kac formula [Azaïs and Wschebor (2008), Adler and Taylor (2007)] to count the expected number of such points.

Neither of these approaches strictly require convexity of the cone generated by the parameter set M. Rather, they depend on a notion of local or infinitesimal convexity referred to as positive reach [Federer (1959)]. Hence, the parameter sets may have finite critical radius. They are both approaches used to form an approximation

$$\mathbb{P}\Big(\sup_{v\in M} f(v) > u\Big) \cong \sum_{j=0}^{\infty} \mathcal{L}_j(M)\rho_j(u) \tag{41}$$

for some centered, smooth, Gaussian process f having constant variance 1 on a (possibly stratified) manifold M. In the volume of tubes approach, the critical radius appears in a natural way and enters into an estimate of the error of the volume of tubes approach. In both approaches, though it is clearer in the expected Euler characteristic approach, the spherical critical radius is in fact the spherical critical radius of $\Psi(M)$ where $\Psi : M \to S(H)$ where $S(H)$ is the unit sphere in H, the reproducing kernel Hilbert space of f. Either approach yields roughly the same

estimate of error: for u large enough

$$\liminf_{u\to\infty} -\frac{2}{u^2}\left|\mathbb{P}\Big(\sup_{v\in M} f(v) > u\Big) - \sum_{j=0}^{\infty} \mathcal{L}_j(M)\rho_j(u)\right| = 1 + \tan^2\big(r_c(M)\big). \tag{42}$$

The above says that the *relative error* in the approximation is exponentially small whenever $r_c(M) > 0$.

The critical radius of M in the expected Euler characteristic approach arises in terms of a functional of the original process f. Specifically, if we assume M is a manifold without boundary, then define for each $x \neq y$ the process introduced in Taylor, Takemura and Adler (2005)

$$f^x(y) = \frac{f(y) - \mathbb{E}(f(y)|f(x), \nabla f(x))}{1 - \mathbb{E}(f(x)\cdot f(y))}. \tag{43}$$

Then,

$$\cot^2\big(r_c(M)\big) = \sup_{x\neq y} \mathbb{E}\big(f^x(y)^2\big). \tag{44}$$

Hence, the critical radius depends in an explicit way on the covariance function of the process f.

As mentioned previously, the argument related to derivation of (44) in the smooth case led directly to the exponential limit in (38). Such a connection suggests a relation between the distribution of the maxima of smooth Gaussian processes, specifically the spacings of the extreme values, can be used to derive weak convergence results for high-dimensional inference. This is ongoing work.

5. Conclusion

We have described what we call the two basic tools of the geometry of least squares that are just as relevant as when Gauss and Legendre disputed their original discovery over 200 years ago. While these tools are not the most technically sophisticated tools, ceding that ground to exponential families for the canonical model (9) and empirical processes for the fluctuation theory in Section 3, they nevertheless provide guiding principles for these more precise tools. We would argue that the Gaussian picture provided by the geometry of least squares, gets much of the picture correct under sufficient moment conditions. For heavier tailed results, of course much of Section 3 would have to be reframed and Section 4 paints quite a different picture [Adler, Samorodnitsky and Taylor (2010, 2013)].

As Bernoulli found himself with just a law of large numbers, the field of statistics is roughly at this same stage in high-dimensional inference. We are hopeful that the geometry of least squares will eventually guide the field to weak convergence results in high-dimensional inference for the canonical problem (9).

Acknowledgement

Supported in part by NSF Grant DMS-12-08857 and AFOSR Grant 113039.

References

Adler, R.J., Samorodnitsky, G. and Taylor, J.E. (2010). Excursion sets of three classes of stable random fields. *Adv. in Appl. Probab.* **42** 293–318. MR2675103

Adler, R.J., Samorodnitsky, G. and Taylor, J.E. (2013). High level excursion set geometry for non-Gaussian infinitely divisible random fields. *Ann. Probab.* **41** 134–169.

Adler, R.J. and Taylor, J.E. (2007). *Random Fields and Geometry*. *Springer Monographs in Mathematics*. New York: Springer. MR2319516

Amari, S.I. and Nagaoka, H. (2000). *Methods of Information Geometry*. *Translations of Mathematical Monographs* **191**. Providence, RI: Amer. Math. Soc. Translated from the 1993 Japanese original by Daishi Harada. MR1800071

Azaïs, J.M. and Wschebor, M. (2008). A general expression for the distribution of the maximum of a Gaussian field and the approximation of the tail. *Stochastic Process. Appl.* **118** 1190–1218. MR2428714

Becker, S., Bobin, J. and Candès, E.J. (2011). NESTA: A fast and accurate first-order method for sparse recovery. *SIAM J. Imaging Sci.* **4** 1–39. MR2765668

Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

Bien, J., Taylor, J. and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Ann. Statist.* To appear. Available at http://arxiv.org/abs/1205.5050.

Boyd, S., Parikh, N. and Chu, E. (2011). *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Hanover: Now Publishers.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge: Cambridge Univ. Press. MR2061575

Bühlmann, P. (2012). Statistical significance in high-dimensional linear models. Available at http://arxiv.org/abs/1202.1377.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Heidelberg: Springer. MR2807761

Candès, E.J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240

Candès, E.J., Romberg, J. and Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory* **52** 489–509. MR2236170

Donoho, D.L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52** 1289–1306. MR2241189

Donoho, D.L., Maleki, A. and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.

Efron, B. (1978). The geometry of exponential families. *Ann. Statist.* **6** 362–376. MR0471152

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. With discussion, and a rejoinder by the authors. MR2060166

Federer, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.* **93** 418–491. MR0110078

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

Friedman, J., Hastie, T. and Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. R package version 1.1-3.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332. MR2415737

Jenatton, R., Mairal, J., Obozinski, G. and Bach, F. (2011). Proximal methods for hierarchical sparse coding. *J. Mach. Learn. Res.* **12** 2297–2334. MR2825428

Laber, E.B. and Murphy, S.A. (2011). Adaptive confidence intervals for the test error in classification. *J. Amer. Statist. Assoc.* **106** 904–913. MR2894746

Lockhart, R., Taylor, J., Tibshirani, R. and Tibshirani, R. (2013). A significance test for the lasso. Available at http://arxiv.org/abs/1301.7161.

Mazumder, R., Hastie, T. and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11** 2287–2322. MR2719857

Meinshausen, N. (2012). Sign-constrained least squares estimation for high-dimensional regression. Available at http://arxiv.org/abs/1202.0889.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473. MR2758523

Minnier, J., Tian, L. and Cai, T. (2011). A perturbation method for inference on regularized regression estimates. *J. Amer. Statist. Assoc.* **106** 1371–1382. MR2896842

Negahban, S.N., Ravikumar, P., Wainwright, M.J. and Yu, B. (2012). A unified framework for high-dimensional analysis of MM-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Math. Program.* **103** 127–152. MR2166537

Obozinski, G., Jacob, L. and Vert, J.P. (2011). Group lasso with overlaps: The latent group lasso approach. Available at http://arxiv.org/abs/1110.0413.

Obozinski, G., Wainwright, M.J. and Jordan, M.I. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Statist.* **39** 1–47. MR2797839

Schneider, R. (1993). *Convex Bodies*: *The Brunn–Minkowski Theory*. *Encyclopedia of Mathematics and Its Applications* **44**. Cambridge: Cambridge Univ. Press. MR1216521

Siegmund, D. and Zhang, H. (1993). The expected number of local maxima of a random field and the volume of tubes. *Ann. Statist.* **21** 1948–1966. MR1245775

Stein, C.M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

Stigler, S.M. (1981). Gauss and the invention of least squares. *Ann. Statist.* **9** 465–474. MR0615423

Sun, J. (1993). Tail probabilities of the maxima of Gaussian random fields. *Ann. Probab.* **21** 34–71. MR1207215

Takemura, A. and Kuriki, S. (1997). Weights of $\overline{\chi}^2$ distribution for smooth or piecewise smooth cone alternatives. *Ann. Statist.* **25** 2368–2387. MR1604465

Takemura, A. and Kuriki, S. (2002). On the equivalence of the tube and Euler characteristic methods for the distribution of the maximum of Gaussian fields over piecewise smooth domains. *Ann. Appl. Probab.* **12** 768–796. MR1910648

Taylor, J.E. (2006). A Gaussian kinematic formula. *Ann. Probab.* **34** 122–158. MR2206344

Taylor, J., Takemura, A. and Adler, R.J. (2005). Validity of the expected Euler characteristic heuristic. *Ann. Probab.* **33** 1362–1396. MR2150192

Taylor, J.E. and Tibshirani, R.J. (2013). Estimation error bounds for convex problems with geometrically decomposable penalties. Unpublished manuscript.

Taylor, J.E. and Vadlamani, S. (2013). Random fields and the geometry of Wiener space. *Ann. Probab.* To appear. Available at http://arxiv.org/abs/1105.3839.

Taylor, J.E. and Worsley, K.J. (2007). Detecting sparse signals in random fields, with an application to brain mapping. *J. Amer. Statist. Assoc.* **102** 913–928. MR2354405

Taylor, J.E. and Worsley, K.J. (2008). Random fields of multivariate test statistics, with applications to shape analysis. *Ann. Statist.* **36** 1–27. MR2387962

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

Tibshirani, R.J. (2012). The lasso problem and uniqueness. Available at http://arxiv.org/abs/1206.0313.

Tibshirani, R.J. and Taylor, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205

Tibshirani, R.J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *Ann. Statist.* **40** 1198–1232. MR2985948

Tseng, P. (2013). On accelerated proximal gradient methods for convex-concave optimization submitted to siam. *J. Optim.* To appear.

Tsirel'son, B.S. (1982). A geometric approach to maximum likelihood estimation for an infinite-dimensional Gaussian location. I. *Teor. Veroyatn. Primen.* **27** 388–395. MR0657940

Vitale, R.A. (2001). Intrinsic volumes and Gaussian processes. *Adv. in Appl. Probab.* **33** 354–364. MR1842297

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689

Weyl, H. (1939). On the volume of tubes. *Amer. J. Math.* **61** 461–472. MR1507388

Worsley, K.J. (1995). Boundary corrections for the expected Euler characteristic of excursion sets of random fields, with an application to astrophysics. *Adv. in Appl. Probab.* **27** 943–959. MR1358902

Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J. and Evans, A.C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping* **4** 58–73.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574