Misfit Measures and Statistical Inconsistency in Linear Inverse Problems

Christopher R. Genovese
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
genovese@stat.cmu.edu

Philip B. Stark
Department of Statistics
University of California
Berkeley CA 94720-3860
stark@stat.berkeley.edu

Model Linear Inverse Problem

Seek to recover a model x, an element of a subset $\mathcal{C}$ of the Hilbert space $\mathcal{X}$, from data $\underline{\delta}$ linearly related to x, but contaminated by observational noise $\underline{\epsilon}$:

$$\begin{aligned} \underline{\delta}_n &= K_n x + \underline{\epsilon}_n \in \mathbf{R}^n, \\ x &\in \mathcal{C} \subset \mathcal{X}, \text{ a Hilbert space,} \\ K_n x &= (\langle\, k_1 \,|\, x \,\rangle, \langle\, k_2 \,|\, x \,\rangle, \ldots, \langle\, k_n \,|\, x \,\rangle), \ \{k_j\}_{j=1}^n \subset \mathcal{X} \\ \{\epsilon_j\}_{j=1}^n &\text{ iid } N(0,1). \end{aligned}$$

$\langle\, \cdot \,|\, \cdot \,\rangle$ is the inner product on $\mathcal{X}$, and $\|y\| = \sqrt{\langle\, y \,|\, y \,\rangle}$.

$\|\underline{\gamma}_n\|_{p,n}$ = usual n-dimensional ℓ_p-norm.

For infinite-dimensional $\underline{\gamma}$, $\|\underline{\gamma}\|_{p,n} = \|\underline{\gamma}_n\|_{p,n}$.

Assume $\mathcal{C}$ has at least 2 elements (otherwise, we know x already).

For any subset $\mathcal{S}$ of a metric space, if $|u-v|$ is the distance between u and v,

$$\begin{aligned} \text{diam}(\mathcal{S}) &= \sup_{u,v \in \mathcal{S}} |u - v|, \\ \text{diam}(\emptyset) &= 0. \end{aligned}$$

$\chi_{p,n,\alpha}$ is $1-\alpha$ quantile of p-norm of n iid $N(0,1)$ variables:

$$\Pr\{\|\underline{\epsilon}_n\|_{p,n} \leq \chi_{p,n,\alpha}\} = 1-\alpha.$$

$$\mathcal{D}_n = \mathcal{D}_{p,n,\alpha} \equiv \{y \in \mathcal{X} : \|K_n y - \underline{\delta}_n\|_{p,n} \leq \chi_{p,n,\alpha}\}$$

$$\Pr_x\{\mathcal{D}_{p,n,\alpha} \ni x\} \geq 1-\alpha.$$

Since $x \in \mathcal{C}$, $\Pr_x\{\mathcal{C} \cap \mathcal{D}_n \ni x\} \geq 1-\alpha$.

Many "inversion" techniques use $\mathcal{D}_n$ and $\mathcal{C} \cap \mathcal{D}_n$.

Example. Minimum-norm estimate (MNE) of x:

$$\hat{x}(\underline{\delta}_n) = \arg \min_{y \in \mathcal{C} \cap \mathcal{D}_n} \|y\|, \text{ whenever } \mathcal{C} \cap \mathcal{D}_n \neq \emptyset.$$

This is a common regularization scheme in geophysics, sometimes called "Occam's Inversion."

Corresponds to a particular choice of regularization parameter in Tichonov regularization.

• C. Constable and R. Parker. Deconvolution of long-core paleomagnetic measurements—spline therapy for the linear problem. *Geophys. J. Int.*, 104:435–468, 1991.

• S.C. Constable, R.L. Parker, and C.G. Constable. Occam's inversion: A practical algorithm for generating smooth models from electromagnetic sounding data. *Geophys.*, 52:289–300, 1987.

• C. deGroot-Hedlin and S. Constable. Occam's inversion to generate smooth, two-dimensional models from magnetotelluric data. *Geophysics*, 55:1613–1624, 1990.

• M. Henry, J.A. Orcutt, and R.L. Parker. A new method for slant stacking refraction data. *Geophys. Res. Lett.*, 7:1073–1076, 1980.

• R.L. Parker. *Geophysical Inverse Theory.* Princeton University Press, Princeton, NJ, 1994.

• L. Shure, R.L. Parker, and G.E. Backus. Harmonic splines for geomagnetic modeling. *Phys. Earth Planet. Inter.*, 28:215–229, 1982.

• L. Shure, R.L. Parker, and R.A. Langel. A preliminary harmonic spline model from Magsat data. *J. Geophys. Res.*, 90:11,505–11,512, 1985.

• P.B. Stark and R.L. Parker. Smooth profiles from $\tau(p)$ and $X(p)$ data. *Geophys. J. R. Astron. Soc.*, 89:997–1010, 1987.

Since $\Pr\{\mathcal{C} \cap \mathcal{D}_n \ni x\} \geq 1 - \alpha$, inequalities satisfied by all $y \in \mathcal{C} \cap \mathcal{D}_n$ are satisfied by x, with confidence $\geq 1 - \alpha$.

Define

$$F^- \equiv \inf_{y \in \mathcal{C} \cap \mathcal{D}_n} F[y]$$

and

$$F^+ \equiv \sup_{y \in \mathcal{C} \cap \mathcal{D}_n} F[y].$$

$[F^-, F^+]$ is a $1 - \alpha$ confidence interval for $F[x]$.

The confidence level is simultaneous for arbitrarily many F.

• G.E. Backus. Confidence set inference with a prior quadratic bound. *Geophys. J.*, 97:119–150, 1989.

• G.E. Backus. Trimming and procrastination as inversion techniques. *Phys. Earth Planet. Inter.,* 98:101–142, 1996.

• J.D. Garmany, J.A. Orcutt, and R.L. Parker. Travel-time inversion: a geometrical approach. *J. Geophys. Res.*, 84:3615–3622, 1979.

• J.R. Grasso, M. Cuer, and G. Pascal. Use of two inverse techniques. Application to a local structure in the New Hebrides island arc. *Geophys. J. R. Astron. Soc.*, 75:437–472, 1983.

• S.P. Huestis and M.E. Ander. IDB2—a FORTRAN program for computing extremal bounds in gravity data interpretation. *Geophysics*, 48:999–1010, 1983.

• H.O. Johnson, D.C. Agnew, and K. Hudnut. Extremal bounds on earthquake movement from geodetic data - application to the Landers earthquake. *Bull. Seis. Soc. Am.*, 84:660–667, 1994.

• S.W. Lang. Bounds from noisy linear measurements. *IEEE Trans. Info. Th.*, IT-31:498–508, 1985.

• T.L. Marzetta and S.W. Lang. Power spectral density bounds. *IEEE Trans. Info. Th.*, IT-30:117–122, 1983.

• M. McNutt and L. Royden. Extremal bounds on geotherms in eroding mountain belts from metamorphic pressure-temperature conditions. *Geophys. J. R. Astron. Soc.*, 88:81–95, 1987.

• D.W. Oldenburg. Funnel functions in linear and nonlinear appraisal. *J. Geophys. Res.*, 88:7387–7398, 1983.

• J.A. Orcutt. Joint linear, extremal inversion of seismic kinematic data. *J. Geophys. Res.*, 85:2649–2660, 1980.

• R.L. Parker and M.A. Zumberge. An analysis of geophysical experiments to test Newton's law of gravity. *Nature*, 342:39–31, 1989.

• J.E. Pierce and B.W. Rust. Constrained least squares interval estimation. *SIAM J. Sci. Stat. Comput.*, 6:670–683, 1985.

• P.C. Sabatier. Positivity constraints in linear inverse problems I. General theory. *Geophys. J. R. Astron. Soc.*, 48:415–441, 1977.

• P.C. Sabatier. Positivity constraints in linear inverse problems II. Applications. *Geophys.*

J. R. Astron. Soc., 48:443–459, 1977.

• C. Safon, G. Vasseur, and M. Cuer. Some applications of linear programming to the inverse gravity problem. *Geophysics*, 42:1215–1229, 1977.

• M. Schlax and D.W. Oldenburg. Age bounds from lead isotope data. *Earth. Planet. Sci. Lett.*, 68:413–421, 1984.

• P.B. Stark. Inference in infinite-dimensional inverse problems: Discretization and duality. *J. Geophys. Res.*, 97:14,055–14,082, 1992.

• P.B. Stark and R.L. Parker. Velocity bounds from statistical estimates of $\tau(p)$ and $X(p)$. *J. Geophys. Res.*, 92:2713–2719, 1987.

• P.B. Stark, R.L. Parker, G. Masters, and J.A. Orcutt. Strict bounds on seismic velocity in the spherical Earth. *J. Geophys. Res.*, 91:13,892–13,902, 1986.

• D.W. Vasco. Bounding seismic velocities using a tomographic method. *Geophysics*, 56, 1991.

• M.A. Zumberge, M.E. Ander, T. V. Lautzenhiser, R. L. Parker, and fourteen others. The greenland gravitational constant experiment. *J. Geophys. Res.*, B10:15,483–15,502, 1990.

$$\frac{\chi_{p,n,\alpha}}{n^{1/2}} = O(1).$$

Allowable misfit grows as $\sqrt{n}$. Unless data image $K_n x$ of x grows even faster with n, with high probability the set $\mathcal{D}_n$ of models that fit the data adequately will eventually contain 0.

The MNE will then be zero, so the error of the MNE will be $\|x\|$.

We might as well not have collected the data.

Need $\|K_n\|$ to grow faster than $n^{1/2}$.

If the components of K_n are orthonormal, $\|K_n\| \to 1$.

Consistency

An estimator $T(\underline{\delta}_n)$ is *consistent over* $\mathcal{C}$ if for every $y \in \mathcal{C}$ and $\forall \gamma > 0$,

$$\lim_{n\to\infty} \Pr_y\{|T(\underline{\delta}_n) - y| > \gamma\} = 0.$$

$\mathcal{I}(\underline{\delta}_n)$ is a $1-\alpha$ *confidence interval* (CI) for $F[x]$ if $\forall y \in \mathcal{C}$ and $\forall n$,

$$\Pr_y\{\mathcal{I}(\underline{\delta}_n) \ni F[y]\} \geq 1 - \alpha.$$

$\mathcal{I}$ is *consistent over* $\mathcal{C}$ if, in addition, $\forall \gamma > 0$,

$$\lim_{n\to\infty} \Pr_y\{\text{diam}(\mathcal{I}(\underline{\delta}_n)) > \gamma\} = 0.$$

$\mathcal{G}(\underline{\delta}_n)$ is a $1-\alpha$ *confidence set* (CS) for x if $\forall y \in \mathcal{C}$ and $\forall n$,

$$\Pr_y\{\mathcal{G}(\underline{\delta}_n) \ni y\} \geq 1 - \alpha.$$

$\mathcal{G}$ is *consistent* if, in addition, $\forall \gamma > 0$,

$$\lim_{n\to\infty} \Pr_y\{\text{diam}(\mathcal{G}(\underline{\delta}_n)) > \gamma\} = 0.$$

.

Consistency is a reasonable minimal requirement: estimate should improve (confidence set should shrink) as $n \uparrow$.

Main Results

• unless the data are highly redundant relative to $\mathcal{C}$, in a precise sense, no estimator or CS can be consistent

• in some additional cases, $\mathcal{C} \cap \mathcal{D}_n$ is an inconsistent CS and the MNE over $\mathcal{C} \cap \mathcal{D}_n$ is an inconsistent estimator

• CI for $F[x]$ derived by minimizing and maximizing $F[y]$ over $\mathcal{C} \cap \mathcal{D}_n$ can be inconsistent

• in some problems in which a nontrivial constraint set $\mathcal{C}$ for the unknown function x is available, using a chi-squared measure of misfit to selected averages of the data yields consistent CS, CI, and MNE

• when some of the conditions required for the previous result fail, it can still happen that using a chi-squared measure of misfit to selected averages of the data yields consistent MNE and CI for finite collections of linear functionals of x.

Apology: lots of definitions, some results obvious.

Inconsistency in Problems with Direct Data

Seek to estimate an n-vector of parameters $\underline{\theta}_n$ from observations $\underline{\delta}_n$ corrupted by a vector $\underline{\epsilon}_n$ of iid standard Gaussian errors:

$$\underline{\delta}_n = \underline{\theta}_n + \underline{\epsilon}_n.$$

First n components of the infinite-dimensional vectors $\underline{\delta}$, $\underline{\theta}$, and $\underline{\epsilon}$.

Study the problem as $n \to \infty$.

Suppose *a priori* that $\underline{\theta} \in \mathcal{C} \subset \mathbf{R}^\infty$.

Prototype for some inverse problems, but we directly observe noisy samples of what we want to know.

Theorem 1 *Suppose $\mathcal{C}$ contains an element $\underline{\psi} \neq \underline{\theta}$ such that $\|\underline{\psi} - \underline{\theta}\|_2 < \infty$, and let $p \geq 1$. No estimator of $\underline{\theta}$ is consistent over $\mathcal{C}$ in ℓ_p norm, and no $1 - \alpha$ CS for x is consistent in ℓ_p norm.*

For there to be a possibility of estimating $\underline{\theta}$ consistently in ℓ_p norm, $\mathcal{C}$ must either contain only $\underline{\theta}$ (which we have assumed is false), or be very strange.

E.g., if $\mathcal{X}$ is ℓ_2, there is no consistent estimator or confidence set for $\underline{\theta}$.

Suppose $\{k_j\}_{j=1}^{\infty}$ are the elements of an orthonormal basis for $\mathcal{X}$.

The jth datum is

$$\delta_j = \langle\, k_j \,|\, x \,\rangle + \epsilon_j,$$

where $\{\epsilon_j\}_{j=1}^{\infty}$ are iid $N(0,1)$.

Want to recover x from these noisy, generalized Fourier coefficients. Since $\{k_j\}$ are orthonormal, by Parseval,

$$\|y\| = \sqrt{\sum_{j=1}^{\infty} \langle\, k_j \,|\, y \,\rangle^2}.$$

Identify $\theta_j = \langle\, k_j \,|\, x \,\rangle$: two-norms in data space and in model space are identical

$$\|y\| = \|\underline{\theta}\|_2 = \lim_{n\to\infty} \|\underline{\theta}\|_{2,n}.$$

Estimating x is equivalent to estimating Kx.

Theorem 1 thus yields

Corollary 2 *There is neither a consistent CS for x nor a consistent estimator of x. In particular, $\mathcal{C} \cap \mathcal{D}_n$ is inconsistent, and MNE over $\mathcal{C} \cap \mathcal{D}_n$ is inconsistent in the L_2 norm on $\mathcal{X}$.*

Theorem 3 *Suppose $w \in \mathcal{X}$. As $n \to \infty$, the length of the CI for $\langle w \,|\, x \rangle$ derived by minimizing and maximizing $\langle w \,|\, y \rangle$ over $y \in \mathcal{C} \cap \mathcal{D}_n$ has a non-zero probability of being equal to the diameter of the one-dimensional projection of $\mathcal{C}$ onto the subspace spanned by w. In particular, if*

$$\mathcal{C} = \{y \in \mathcal{X} : \|y\| \leq C\},$$

the probability that the length of the CI is $2C\|w\|$ converges to a positive value.

Asymptotically, using the data in this way may tell us nothing more than we knew *a priori*.

As before, but data mapping functionals $\{k_j\}$ are bounded and linearly independent, not necessarily orthogonal.

Linear independence implies that each observation contains at least some new information about x, so we do not get repeated observations of exactly the same properties of x.

Corollary 4 *Suppose $\{k_j\}$ is a linearly independent bounded subset of $\mathcal{X}$. If $\mathcal{C}$ contains $y \neq x$ s.t. $\|K(x-y)\| < \infty$, no estimator or CS is consistent over $\mathcal{C}$.*

Theorem 5 *If $\mathcal{C}$ contains 0 and $x \neq 0$, and $\{k_j\}$ is a linearly independent, bounded subset of $\mathcal{X}$ s.t.*

$$\limsup_{n\to\infty} \frac{1}{\sqrt{n}} \|K_n x\|_{2,n}^2 = 0, \text{ then}$$

A. *The CS $\mathcal{C} \cap \mathcal{D}_n$ is inconsistent.*

B. *If $w \in \mathcal{X}$, the CI*

$$[\inf_{y\in\mathcal{C}\cap\mathcal{D}_n} \langle\, w \,|\, y \,\rangle, \sup_{y\in\mathcal{C}\cap\mathcal{D}_n} \langle\, w \,|\, y \,\rangle\,] \tag{1}$$

for $\langle\, w \,|\, x \,\rangle$ is inconsistent.

C. *The MNE of x over $\mathcal{D}_n$ is inconsistent in the norm of $\mathcal{X}$.*

If $\{k_j\}$ is a generalized Fourier basis and the constraint set $\mathcal{C}$ is a norm ball with positive radius in the model space $\mathcal{X}$, the conditions are met: the norm in condition (5) must be finite; after dividing it by $\sqrt{n}$ it must converge to zero.

Condition (5) limits the redundancy in the measurements. Two stronger conditions that imply (5) are

(i) $\limsup_{n\to\infty} n^{-1/2} \|K_n z\|_{2,n}^2 = 0$ for all $z \in \mathcal{C}$, and

(ii) $\limsup_{n\to\infty} n^{-1/2} \|K_n\|^2 = 0$, where $\|K_n\|$ is the operator norm of the data mapping at stage n.

Consistency of the CS $\mathcal{C} \cap \mathcal{D}_n$ depends on the prior information.

Can make a much more general statement about the behavior of $\mathcal{C} \cap \mathcal{D}_n$ when $\mathcal{C}$ has non-empty interior in the topology of the model space $\mathcal{X}$.

Theorem 6 *Suppose $\mathcal{C}$ has non-empty interior in the topology of $\mathcal{X}$. Then the CS $\mathcal{C} \cap \mathcal{D}_n$ is inconsistent whenever x lies in the interior of $\mathcal{C}$.*

This does not imply that MNE is necessarily inconsistent.

Similarly, projection of x onto finite-dimensional subspaces (finite sets of linear functionals of x) still might be estimated consistently.

Averaging the data

• diam($\mathcal{D}_n$) does not shrink because each new datum has a new error: must increase allowable misfit to maintain $1 - \alpha$ confidence.

• Had the noise variance σ^2, the radius of the ℓ_p misfit ball would be multiplied by σ.

• Can simulate a decreasing noise level by averaging large groups of data as $n \uparrow$.

• In many problems, averaging entails irrevocable loss of information about x.

• If many data are essentially redundant given that $x \in \mathcal{C}$, can average without losing information.

Similar approaches long used in probability density estimation and spectrum estimation.

In kernel density estimates, to attain consistency, the width of the kernel must decrease as n increases, but sufficiently slowly that an increasing number of points contribute significantly to the estimate at any point where the density does not vanish.

Raw periodogram estimate of spectral density is inconsistent, but if the periodogram is averaged in bins in such a way that the number of points in each bin goes to infinity, but the width of each bin shrinks to zero, the resulting estimate is consistent.

Tradeoff between bias and variance: making the kernel or bins wider increases the number of data contributing to the estimate at each point, which decreases the variance of the estimate, but a wider kernel or bin involves more of the function, and averaging neighboring values together biases the estimate.

Both the bias and the variance must go to zero to get consistency, so the bins must get narrower and narrower, but must still contain more and more observations.

Similar ideas show that basing estimates or CS on suitable data averages can yield consistency. Need to average larger and larger groups of observations, but each group must measure increasingly similar properties of the model or bias won't go away.

In spectrum estimation and probability density estimation, there is a natural order to the data: sensible to average neighbors.

In more general problems, replace idea of neighboring points by a more abstract notion of continguity: how much the functionals can differ when applied to distinct elements of $\mathcal{C}$.

If they cannot differ much, averaging them introduces only a small bias.

Must ensure that can choose subsets of the data to average so as to lose no information essential to identifying which element of the constraint set $\mathcal{C}$ the true model x is.

Requires the data to be highly redundant relative to $\mathcal{C}$: essentially every property that differs among elements of $\mathcal{C}$ must be measured infinitely many times.

This is not met in the generalized Fourier reconstruction problem. Any averaging at all prevents us from being able to identify

x uniquely. Can get consistent estimates in those cases in which averaging does not cost us too much information about x.

Example to hold in mind: estimating a bandlimited function x from its convolution with an analytic kernel, sampled at a set of points that grow increasingly dense as $n \to \infty$, with noise added.

If we knew the convolution on an open set, could deconvolve and reconstruct x everywhere. If two bandlimited functions have the same convolution at any dense set of points, they are identical—the measurements *separate* the set of bandlimited functions. Don't need observations to grow increasingly dense everywhere, just on some open set. Isolated observations in some places are not essential to reconstructing x.

As in spectrum estimation, averaging neighboring observations is attractive, because they sample approximately the same part of x. Need bins to shrink so that the bias vanishes, but shrink sufficiently slowly that the number of observations in each bin grows, so the variance vanishes too.

Asymptotic modulus of continuity of the problem $(K, \mathcal{C})$:

$$\omega_{\mathcal{C}}(\nu) = \sup\left\{\|y - z\| \,:\, y, z \in \mathcal{C}\,;\, \limsup_n |K_n(y - z)|_n \leq \nu\right\}.$$

Measures how well $\{k_j\}_{j=1}^n$ constrain objects in $\mathcal{C}$ as more and more of them are observed.

$\{k_j\}_{j=1}^n$ has the *asymptotic separation property* over $\mathcal{C}$ if $\omega_{\mathcal{C}}(\nu) \to 0$ as $\nu \to 0$.

$\mathcal{C}$-distance between two linear functionals k_i and k_j:

$$|k_i - k_j|_{\mathcal{C}} \equiv \sup_{y,z \in \mathcal{C}} |\langle\, k_i \,|\, y - z \,\rangle - \langle\, k_j \,|\, y - z \,\rangle|.$$

Replaces our intuitive notion of the neighborhood of a point in more abstract problems.

A point $k_j \in \{k_j\}$ is *densely covered* if every $\mathcal{C}$-neighborhood of k_j contains at least one other member of $\{k_j\}$; i.e., for every $\gamma > 0$,

$$\#\{i \neq j : |k_j - k_i|_{\mathcal{C}} < \gamma\} \geq 1.$$

Let $\{k_j\}^{\mathrm{d}}$ be the set of densely covered elements of $\{k_j\}$.

If some element of $\{k_j\}$ that is not densely covered is essential to recovering x, cannot take smaller and smaller neighborhoods and still have more and more data in each neighborhood as n grows, so can't drive the bias to zero.

$\{k_j\}$ has the *finite coverage property* if for all $\xi > 0$, there is an n_0 such that for all $n \geq n_0$ and for all $k \in \{k_j\}^{\mathrm{d}}$

$$\inf_{1 \leq j \leq n} |k_j - k|_{\mathcal{C}} \leq \xi.$$

For the deconvolution problem, this is equivalent to requiring that for any $\xi > 0$, there is some finite stage n by which time we have measurements within a distance ξ of every measurement we will ever get. Satisfied, e.g., if data taken on a dyadic grid in a bounded interval.

Finite coverage is *regular* if there exists $\eta > 0$ such that for all $\xi > 0$, whenever we choose an n_0 as above,

$$\limsup_{n \geq n_0} \frac{\max_{1 \leq j \leq n_0} \#\{1 \leq i \leq n \,:\, |k_j - k_i|_{\mathcal{C}} \leq \xi\}}{\min_{1 \leq j \leq n_0} \#\{1 \leq i \leq n \,:\, |k_j - k_i|_{\mathcal{C}} \leq \xi\}} \leq \eta.$$

Intuitively, rates at which we visit the $\mathcal{C}$-neighborhood of each data functional k_j are comparable.

For the deconvolution problem, says that the order in which we get the data allows us to have samples that grow closer together at about the same pace, they don't "pile up" in some places and remain sparse in others.

$(K, \mathcal{C})$ has the *asymptotic intersection property* if for all $y \in \mathcal{C}$, $\liminf \Pr_y\{\mathcal{C} \cap \mathcal{D}_n \neq \emptyset\} = 1$. Ensures that MNE is almost-surely defined when n is large.

Theorem 7 *If there is a non-empty* $\mathcal{A} \subset \{k_j\}^{\mathrm{d}}$ *s.t.* $\mathcal{A}$ *has the asymptotic separation and regular finite coverage properties over* $\mathcal{C}$*, then there is a way to average subsets of the observations so that using the chi-squared measure of misfit to the data averages yields consistent CS. Furthermore, if* $(K, \mathcal{C})$ *has the asymptotic intersection property, MNE over those CS will be consistent.*

Spirit of the theorem is that if the data mapping asymptotically distinguishes among members of $\mathcal{C}$, and enough data measure essentially the same property of x, can average ever groups of observations (driving the noise level down faster than the number of degrees of freedom goes up) without losing information.

When the asymptotic intersection property does not hold, the definition of MNE and its performance depend explicitly on more subtle properties of $\mathcal{C}$.

Useful to consider what happens when these conditions don't hold.

Let $\mathcal{S}$ be a closed subspace of $\mathcal{X}$, and let $\mathcal{S}^\perp$ be its orthogonal complement.

Let $P_{\mathcal{S}}$ be the orthogonal projection operator onto $\mathcal{S}$, let and $P_{\mathcal{S}^\perp} = I - P_{\mathcal{S}}$ be the projection operator onto $\mathcal{S}^\perp$.

If $\mathcal{F}$ and $\mathcal{G}$ are subsets of $\mathcal{X}$, let $\mathcal{F} - \mathcal{G}$ denote the set $\{y - z : y \in \mathcal{F}, z \in \mathcal{G}\}$.

Suppose $\mathcal{A}_n = \{a_j\}_{j=1}^n$ is the set containing the first n elements of the sequence $(a_j)_{j=1}^\infty$ of members of $\mathcal{X}$.

If $\mathcal{F}$ and $\mathcal{G}$ are any two subsets of $\mathcal{X}$, $\mathcal{F}$ and $\mathcal{G}$ are *asymptotically orthogonal with respect to* $\mathcal{A}$, (written $\mathcal{F} \perp_{\mathcal{A}} \mathcal{G}$), if for every $y \in \mathcal{F}$ and $z \in \mathcal{G}$,

$$\limsup_n n^{-1} \left| \sum_{j=1}^n \langle\, a_j \,|\, y \,\rangle \langle\, a_j \,|\, z \,\rangle \right| = 0.$$

When $\mathcal{A} = \{k_j\}$, write $\mathcal{F} \perp_K \mathcal{G}$.

When the data mappings don't asymptotically separate points of $\mathcal{C}$, the averaging still yields consistent estimates and CS for the projections of x onto appropriate subspaces of $\mathcal{X}$:

Corollary 8 *If there is a non-empty $\mathcal{A} \subset \{k_j\}^{\mathrm{d}}$ such that $\mathcal{A}$ has the regular finite coverage property over $\mathcal{C}$, then averaging subsets of the observations as in Theorem 7 and using the chi-squared measure of misfit to those averages yields consistent CS for $P_{\mathcal{S}}x$ for any closed subspace $\mathcal{S}$ of $\mathcal{X}$ such that $\mathcal{A}$ has the asymptotic separation property over $P_{\mathcal{S}}\mathcal{C}$, and $(P_{\mathcal{S}}\mathcal{C} - P_{\mathcal{S}}\mathcal{C}) \perp_{\mathcal{A}} (P_{\mathcal{S}^\perp}\mathcal{C} - P_{\mathcal{S}^\perp}\mathcal{C})$.*

Moreover, if the asymptotic intersection property holds for $(K, \mathcal{C})$, then the projection of the MNE, $P_{\mathcal{S}}\hat{x}$, is consistent for $P_{\mathcal{S}}x$.

If the effects of elements of a subspace $\mathcal{S}$ on the data are asymptotically orthogonal to the effects of elements of $\mathcal{S}^\perp$ on the data, averaging allows us to recover $P_{\mathcal{S}}x$ consistently, even when we cannot recover the entire object.

If the asymptotic orthogonality does not hold, there are unresolvable tradeoffs with parts of the model we cannot estimate accurately.

Simple example: suppose $\mathcal{C} = \mathcal{X}$ and that $k_j = k_1$, $j = 1, 2, \cdots$. Impossible to determine x by measuring only one of its components, no matter how often, but it is clearly possible to estimate $\langle\, k_1 \,|\, x \,\rangle$, the component of x in the subspace $\mathcal{S}$ spanned by k_1, arbitrarily well.

Application to Geomagnetism

Idealized problem of estimating the scalar potential Ψ of Earth's main magnetic field $\mathbf{B}(\mathbf{r})$ on the sphere $r = a$ (idealization of the core-mantle boundary, CMB) from satellite observations of $\mathbf{B}$ on the surface of the sphere $r = c$, $c > a$.

Magnetic field outside CMB from currents in the core is the gradient of a scalar field Ψ:

$$\mathbf{B} = -\nabla\Psi,$$

where Ψ has the spherical harmonic expansion

$$\Psi(\mathbf{r}) = a \sum_{l=1}^{\infty} (a/r)^{l+1} \sum_{m=-l}^{l} x_{lm}(a) Y_{lm}(\hat{\mathbf{r}}).$$

$\mathbf{r}$ is the position vector with origin at Earth's center, $r = |\mathbf{r}|$ is the Euclidean length of $\mathbf{r}$, $\hat{\mathbf{r}} = \mathbf{r}/|\mathbf{r}|$, and Y_{lm} are spherical harmonics.

Prior information: rest mass of the energy of $\mathbf{B}$ is less than Earth's mass.

$$\mathcal{C} = \{\Phi = a \sum_{l=1}^{\infty} (a/r)^{l+1} \sum_{m=-l}^{l} y_{lm} Y_{lm}(\hat{\mathbf{r}}) : \sum_{l=1}^{\infty} \sum_{m=-l}^{l} q(l)|y_{lm}|^2 \leq 1\},$$

with

$$q_l = (2l+1)(l+1)^{-1}/(2 \times 10^{33} \mathrm{nT}^2),$$

when the units of x_{lm} are nanoTesla (nT).

$\mathcal{X}$ is the Hilbert space of potentials whose sequences of spherical harmonic coefficients (x_{lm}) are square-summable w.r.t. (q_l).

Pretend density of satellite samples asymptotically uniform on $r = c$.

Take $n = 3i$; the n data at a given stage are the three components of $\mathbf{B}(\mathbf{r})$ at $i = n/3$ points on $r = c$.

Let K_n be the mapping from the space of potentials of finite-energy fields to the three components of $\mathbf{B}$ at $n/3$ approximately equally spaced points on $r = c$.

$\mathbf{B}(\mathbf{r}_j)$, $r_j = c$, is related to the spherical harmonic expansion of the potential with coefficients (y_{lm}) on $r = a$ via

$$\mathbf{B}_y(\mathbf{r}_j) = \sum_{l=1}^{\infty} (a/c)^{l+2} \sum_{m=-l}^{l} y_{lm} \nabla [r^{-l-1} Y_{lm}(\hat{\mathbf{r}}_j)]_{r=1}.$$

$\underline{\epsilon}$ consists of

- random measurement errors from instrument noise and uncertainties in satellite orientation
- magnetic fields from all sources exterior to the core

Pretend these combine to yield i.i.d. $N(0, \sigma^2)$ errors, σ known.

Largest error in this approximation is the spatial correlation of the crustal field as observed at satellite altitudes.

• $\mathcal{C}$ is a norm-ball in $\mathcal{X}$, so it has nonempty interior; cannot get consistent CS for Ψ using ℓ_p misfit balls.

• The MNE of Ψ over $\mathcal{D}_n$ is also inconsistent.

• $\{k_j\}$ does not asymptotically separate points of $\mathcal{C}$, so Theorem doesn't say we can recover x using averaged data.

Data averaging does give consistent confidence sets for the projection of Ψ onto the span of any finite number of spherical harmonics: $\mathcal{S} = \text{span}\{Y_{lm}\}_{0 \le l \le l_{\max}, -l \le m \le l}$.

Asymptotic Separation of $\mathcal{C}$ by $\{k_j\}$

Need to show $\omega_{P_{\mathcal{S}}\mathcal{C}}(\nu) \to 0$ as $\nu \to 0$, where

$$\omega_{P_{\mathcal{S}}\mathcal{C}}(\nu) = \sup\{\|y - z\| : y, z \in P_{\mathcal{S}}\mathcal{C}, \limsup_n |K_n(y - z)|_n \leq \nu\}.$$

$\mathcal{C}$ is symmetric, so $\{y - z : y, z \in \mathcal{C}\} = 2\mathcal{C}$, and

$$\omega_{P_{\mathcal{S}}\mathcal{C}}(\nu) \equiv \sup\{\|y\| : y \in 2P_{\mathcal{S}}\mathcal{C} \text{ and } \limsup_n |K_n y|_n \leq \nu\}.$$

Asymptotically equivalent to a linear program.

$$\underline{1} \equiv (1)_{j=1}^{l_{\max}}, \qquad \underline{0} \equiv (0)_{j=1}^{l_{\max}},$$
$$\underline{q}' \equiv (q_l')_{l=1}^{l_{\max}}, \qquad \underline{w} \equiv (w_l)_{l=1}^{l_{\max}},$$

where $w_l = \frac{1}{3}(a/c)^{2(l+2)}(l+1)$.

$$\omega^2_{P_{\mathcal{S}}\mathcal{C}}(\nu) = \sup\{\underline{1} \cdot \underline{p} : \underline{p} \geq \underline{0} \text{ and } \underline{w} \cdot \underline{p} \leq \nu^2 \text{ and } \underline{q}' \cdot \underline{p} \leq 4\}.$$

As $\nu \to 0$, constraint $0 \leq p_l \leq 3\nu^2(a/c)^{-2(l+2)}(l+1)^{-1}$ is eventually stronger in every component than the constraint $\underline{q}' \cdot \underline{p} \leq 4$.

Only one constraint, so by fundamental theorem of LP, optimal solution has only one nonzero element, $p_{l_{\max}}$, and

$$\omega_{P_{\mathcal{S}}\mathcal{C}}(\nu) \to \nu(a/c)^{-l_{\max}-2}(l_{\max}+1)^{-1/2} \to 0 \text{ as } \nu \to 0.$$

Thus $\{k_j\}$ asymptotically separates elements of $P_{\mathcal{S}}\mathcal{C}$.

Regular Finite Coverage

Follows from sampling scheme.

Asymptotic orthogonality of $P_{\mathcal{S}}\mathcal{C}$ and $P_{\mathcal{S}^\perp}\mathcal{C}$ relative to K

Follows from orthonormality of spherical harmonics and asymptotic eigenstructure of K_n.

$\{k_j\}$ is densely covered

The potential is an harmonic function; away from $r = a$, it is analytic, and therefore on $r = c$, it is continuously differentiable. The derivative is uniformly continuous on the compact set $r = c$, and because the data sampling points on $r = c$ grow uniformly closer as $n \to \infty$, for any k_j and any $\gamma > 0$, there is at least one functional $k_i \neq k_j$ whose $\mathcal{C}$-distance from k_j is less than γ.

Conclusions

• When the measurements in an inverse problem are not sufficiently redundant, there is neither a consistent CS for the model nor a consistent estimator.

• In a slightly larger class of problems, CS and MNE estimates based on chi-squared misfit to the data are statistically inconsistent. For example, in linear inverse problems in Hilbert spaces with bounded data functionals, the CS are inconsistent whenever the prior constraint set has nonempty interior.

• Consistent CS and MNE based on the ℓ_p measure of misfit to suitable averages of the data are possible when the observations are sufficiently redundant, given that $x \in \mathcal{C}$.

• Data reduction by fitting to averages of the data can yield substantial computational economies, since the dimension of the problems one needs to solve is much smaller than it would be for the original data set: fractional powers are typical.

Open Questions

• When is the MNE or CS based on ℓ_p misfit consistent over a nontrivial set $\mathcal{C}$ without data averaging?

• In situations in which consistency is possible, how should the data be averaged to yield the best rate of convergence?

• When is the best rate of convergence obtainable by data averaging the optimal rate? (Works for nonparametric regression of Lipschitz functions.)

• For what n does data averaging improve upon fitting to the original data?

• How to characterize the conditions in ways that are easier to verify?

These questions appear to require very specific information about the prior constraint $\mathcal{C}$ and the data mapping K