Statistical Foundations of Data Science and their Applications
A conference in celebration of Jianqing Fan's 60th Birthday

Inducement of population-level sparsity

Heather Battey
Department of Mathematics, Imperial College London

Princeton. May 9, 2023

MOTIVATION

A large number of nuisance parameters*.

A high-dimensional nuisance parameter.

Failure of maximum likelihood theory.

Similar issues arise in Bayesian inference.

*Nuisance parameters are those needed to complete the specification of the probabilistic model but of no direct subject-matter concern.

SPARSITY

- Existence of many zeros or near-zeros.
- Two roles: (1) to aid interpretation; (2) to restrain estimation error.
- This talk: interpretation is central; high-dimensional parameters are nuisance – focus therefore on (2).
- Explore the idea of systematically inducing particular forms of sparsity on population-level quantities.
- Traverse parameterisation space or transformation space with a view to inducing sparsity.

ONE OLD AND THREE NEW EXAMPLES

- RP Parameter orthogonalisation (Cox and Reid, 1987).
- RP Sparsity scales in covariance estimation.
- DT Construction of factorisable transformations.
- DT Inference in high-dimensional regression.

The first two aim to induce sparsity through reparameterisation (RP), the last two via transformations of the data (DT).

Example 1

Parameter orthogonalisation

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc., B*, 49, 1–39.

PARAMETER ORTHOGONALITY

- ψ and λ: interest and nuisance parameters.
- Let $i_{\psi\lambda}(\psi, \lambda)$ denote the corresponding block of the Fisher information matrix. ψ and λ are said to be orthogonal if $i_{\psi\lambda}(\psi, \lambda) = 0$. Global/local.
- Implications: ML for ψ behaves "almost as if" λ was fixed at its true value λ^*.
- "almost as if": if the dimension of λ is fixed then $\widehat{\psi} - \widehat{\psi}_{\lambda^*} = O_p(n^{-1})$.

PARAMETER ORTHOGONALISATION
(Cox and Reid, 1987)

- Starting with a parameterisation (ψ, ϕ), an interest-respecting reparameterisation $(\psi, \lambda(\psi, \phi))$ is chosen to make λ orthogonal to ψ.
- In other words, to induce sparsity on $i_{\psi\lambda}$ (population-level sparsity).
- This is operationalised by solving a system of partial differential equations.
- Some of the modern high-dimensional inference literature implicitly or explicitly assumes $i_{\psi\lambda}$ is sparse without such preliminary manoeuvres.

Example 2

Sparsity-inducing parameterisations for covariance models

Battey, H. S. (2017). Eigen structure of a new class of structured covariance and inverse covariance matrices. *Bernoulli*, 23, 3166–3177.

Rybak, J. and Battey, H. S. (2021). Sparsity induced by covariance transformation: some deterministic and probabilistic results. *Proc. Roy. Soc. Lond. A*, 477.

COVARIANCE MATRICES

- Covariance matrices and their inverses are often nuisance parameters.
- A sparsity assumption allows construction of estimators that are consistent in relevant matrix norms when $p(n)/n \to c > 0$.
- But consistency is only interesting insofar as the assumptions made are satisfied to an adequate order of approximation.
- This motivates a search for parameterisations in which the relevant covariance models are sparse.

AN OPEN PROBLEM

Q*: For a given (relevant) covariance model, not obviously sparse in any domain, can a sparsity-inducing parameterisation be deduced?

A*: . . .

STATISTICAL IMPLICATIONS OF A^*

- Reparameterise to achieve maximal sparsity.
- Seek a more effective and valid statistical analysis on the transformed scale by exploiting the sparsity.
- Transform the conclusions back to the scale of interest.
- Hope, then prove, that the strong statistical properties are preserved after back-transformation (*Biometrika*, 106, 605–617).

Q^*: For a given (relevant) covariance model, not obviously sparse in any domain, can a sparsity-inducing parameterisation be deduced?

A^*: ...

A proof of concept for Q^*: an illustration of the possibility of increasing sparsity through reparameterisation.

EXAMPLE OF A NON-STANDARD PARAMETERISATION

The matrix logarithm L of a covariance matrix Σ is defined as

$$\Sigma = \exp(L) = \sum_{k=0}^{\infty} \frac{1}{k!} L^k.$$

Spectral decomposition:

$$\begin{aligned} \Sigma &= \Gamma \Lambda \Gamma^T, \quad \Lambda \triangleq \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\} \\ L &= \Gamma \Delta \Gamma^T, \quad \Delta \triangleq \mathrm{diag}\{\log(\lambda_1), \ldots, \log(\lambda_p)\}. \end{aligned}$$

The inverse satisfies $\Sigma^{-1} = \exp(-L)$.

WHAT STRUCTURE IS INDUCED ON Σ THROUGH SPARSITY OF L?

$$\begin{aligned}\Sigma,\ \Sigma^{-1} \in \mathcal{V}_p^+(\mathbb{R}) &\triangleq \left\{S \in \mathcal{M}_p(\mathbb{R}) : S = S^T,\ S \succ 0\right\} \quad (\text{open cone})\\ L \in \mathcal{V}_p(\mathbb{R}) &\triangleq \left\{S \in \mathcal{M}_p(\mathbb{R}) : S = S^T\right\} \quad (\text{vector space}).\end{aligned}$$

Natural symmetrised basis for $\mathcal{V}_p(\mathbb{R})$ of the form $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$:

$$\begin{aligned}\mathcal{B}_1 &= \left\{B : B = e_j e_j^T,\ \ j \in [p]\right\}\\ \mathcal{B}_2 &= \left\{B : B = e_j e_k^T + e_k e_j^T,\ \ j,k \in [p],\ j \neq k\right\}.\end{aligned}$$

By contrast, $\mathcal{V}_p^+(\mathbb{R})$ does not possess a basis.

$$L = \sum_{m=1}^{|B|} \alpha_m B_m \ \text{ where } B_1, \ldots, B_{|\mathcal{B}|} \in \mathcal{B}.$$

WHAT STRUCTURE IS INDUCED ON Σ THROUGH SPARSITY OF L?

- Impose sparsity on

$$L = \sum_{m=1}^{|B|} \alpha_m B_m \text{ where } B_1, \ldots, B_{|\mathcal{B}|} \in \mathcal{B}.$$

  through the basis coefficients. Specifically:

$$\alpha = (\alpha_1, \ldots, \alpha_{|\mathcal{B}|}) \text{ satisfies } \|\alpha\|_0 = s^* < p.$$

- The eigenvectors and eigenvalues of Σ inherit substantial structure.

STRUCTURE INDUCED ON THE EIGENVECTORS AND EIGENVALUES OF Σ THROUGH SPARSITY OF L

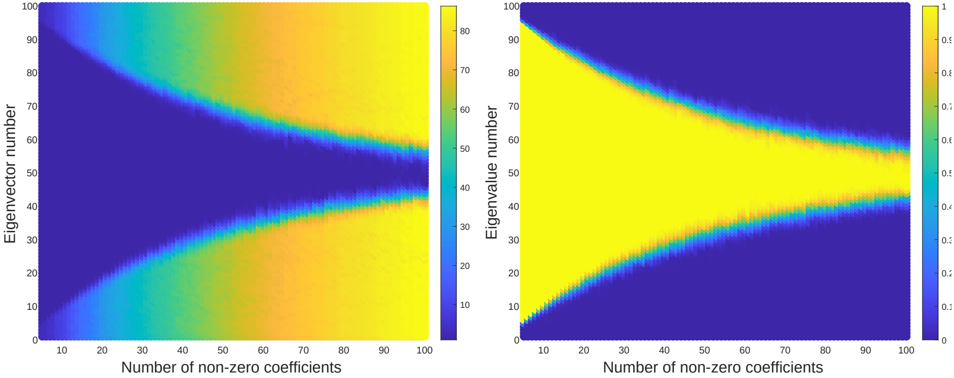


Figure: Simulation average of $\|\gamma_j\|_0$ (left) and $\mathbb{I}\{\lambda_j = 1\}$ (right) for 100 random logarithmically s^*-sparse covariance matrices, plotted against index j of ordered eigenvalues (y-axis) and $s^* \in \{1, \ldots, p\}$ (x-axis) for $p = 100$.

WHAT STRUCTURE IS INDUCED ON Σ THROUGH SPARSITY OF L?

There is a deterministic answer. A random matrix perspective aids interpretation.

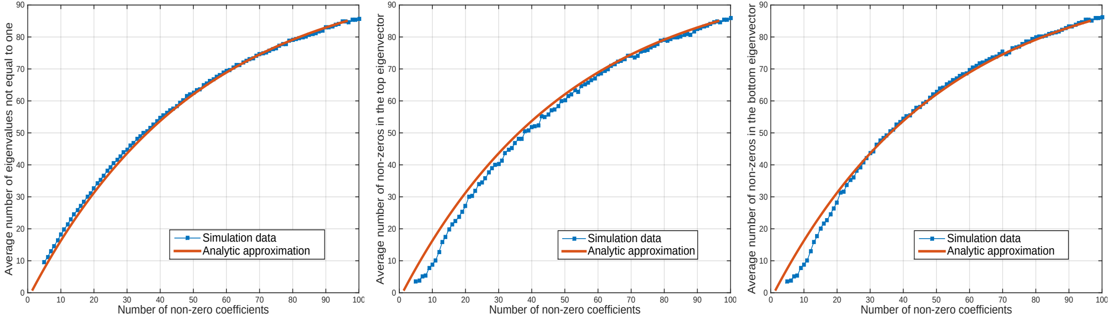
Suppose the support of α is a simple random sample of size s^* from the index set $\{1, \ldots, p(p+1)/2\}$.

- The expected number of non-unit eigenvalues of $\Sigma = \Sigma(\alpha)$ is approximately $d^* < p$, where

$$d^* = \text{root}\left\{ \frac{4p + p(p-1)}{2(p+1)} \left[\log\left(\frac{p}{p-d}\right) - \frac{d}{2p(p-d)} \right] - s^* \right\}.$$

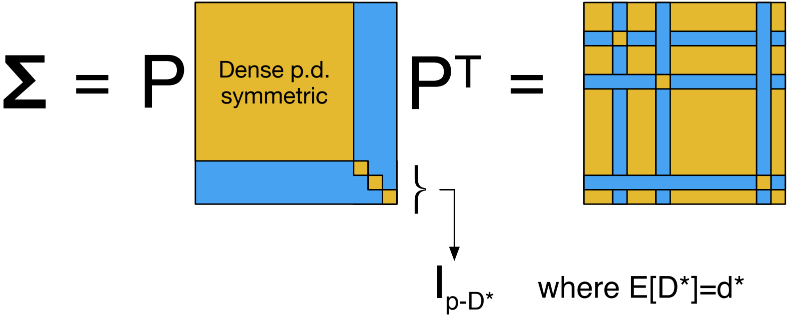
- The corresponding eigenvectors have d^* non-zeros in expectation.
- The other eigenvectors are of the form e_j.

APPROXIMATION ERROR



WHAT STRUCTURE IS INDUCED ON Σ THROUGH SPARSITY OF L?

Suppose the support of α is a simple random sample of size s^* from the index set $\{1, \ldots, p(p+1)/2\}$. The resulting Σ is of the form



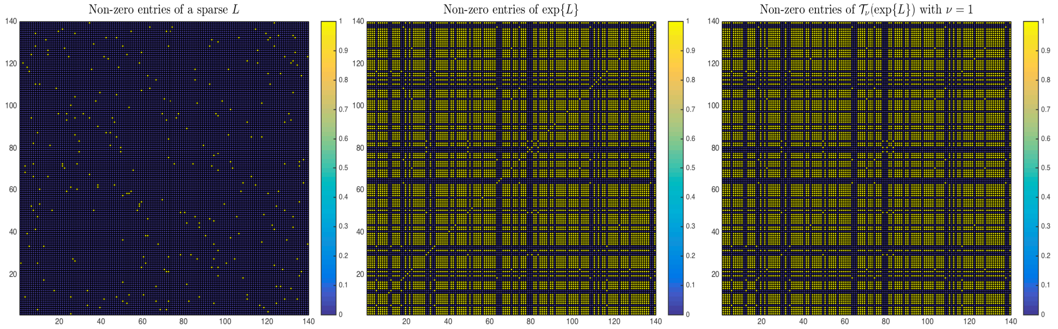
where P is a permutation matrix. The same structure holds for deterministic logarithmically sparse covariance matrices but the dimension of the identity block is less explicit.

WHAT STRUCTURE IS INDUCED ON Σ THROUGH SPARSITY OF L?

Indicator of non-zero entries for:

- Left: one realisation of a random sparse L;
- Centre: the corresponding matrix exponential $\Sigma = \exp(L)$
- Right: the thresholded version $\mathcal{T}(\Sigma) = \{\Sigma_{ij}\mathbb{I}(|\Sigma_{ij}| \geq 1)\}$.

Yellow entries represent non-zeros. Blue entries represent zeros.



A sparse L typically corresponds to an appreciably less sparse Σ.

REMARKS AND OPEN PROBLEMS

- The deterministic version is an if and only if result.
- The same structure holds on Σ^{-1} as on Σ.
- Two other qualitatively similar examples and an encompassing formulation.
- Seek the most appropriate sparsity scale. Analytically? Empirically?
- An empirical approach: parameterise a path through parameterisation space and estimate the sparsity scale, similarly to Box-Cox.

Example 3

Construction of factorisable transformations

Battey, H. S. and Cox, D. R. (2020). High-dimensional nuisance parameters: an example from parametric survival analysis. *Information Geometry*, 3, 119–148.

Battey, H. S., Cox, D. R. and Lee, S. (2023). On partial likelihood and the construction of factorisable transformations. *Information Geometry*, to appear.

SIMPLE MOTIVATING EXAMPLES

- Given n pairs of twins, one twin from each pair is chosen at random to receive a treatment. A response (e.g. recovery time, blood pressure) is measured on the treated twin and control twin and written (T_i, C_i) for $i = 1, \ldots, n$.
- Goal: estimate the treatment effect ψ. There are pair-specific nuisance parameters $\lambda_1, \ldots, \lambda_n$.
- Misleading estimates of ψ are obtained unless problem-specific manoeuvres are applied.
- Eliminate n nuisance parameters: if T_i and C_i are exponentially distributed of rates $\lambda_i + \psi$ and $\lambda_i - \psi$ the conditional density of T_i at t, given $s_i = t_i + c_i$ is

$$\frac{2\psi e^{-2\psi t}}{1 - e^{-2\psi s_i}} \qquad (\text{does not depend on } \lambda_i).$$

- Eliminate n nuisance parameters: if T_i and C_i are exponentially distributed of rates $\lambda_i \psi$ and λ_i / ψ the marginal density of $S_i = T_i / C_i$ at s is

$$\frac{\psi^2}{(1 + \psi^2 s)^2} \qquad (\text{does not depend on } \lambda_i).$$

REMARKS ON THE EXAMPLES

- Easy examples. Inferential separations/group structure.
- Nuisance parameters can be eliminated exactly only in special cases.
- The goal is not to solve specific simple problems, but to reach the correct answer by a seamless application of theory.
- Hope: a general theory that covers exact cases can be applied to more difficult problems to approximately eliminate nuisance parameters.

A SYSTEMATIC ROUTE TO THE ELIMINATION OF NUISANCE PARAMETERS

- Solve a PDE to find a transformation of the data that eliminates the nuisance parameters in cases where an exact solution is available.
- Details omitted. In solving the PDE, population-level sparsity is induced.
- Cox and Reid (1987): operates on the parameter space, reduces the role of nuisance parameters but does not eliminate them.
- Present approach operates on the sample space. Induces a stronger form of sparsity than Cox and Reid (1987).

Example 4

Inference in high-dimensional regression

Battey, H. S. and Reid, N. (2023). On inference in high-dimensional regression. *J. Roy. Statist. Soc., B*, 85, 149-175.

INFERENCE IN HIGH-DIMENSIONAL REGRESSION

- Apply similar ideas in high-dimensional regression problems, i.e. eliminate nuisance parameters to the extent feasible using transformations of the data.
- Inference on regression parameters will be embedded within the inferential framework of confidence sets of models.

INFERENCE IN HIGH-D LINEAR REGRESSION: NOTATION

- n observations from a linear regression model. In matrix notation:

$$Y = X\beta + \varepsilon = x_v\beta_v + X_{-v}\beta_{-v} + \varepsilon.$$

- β is of dimension $p \gg n$ and sparse: $\|\beta\|_0 = s \ll n$.
- Treat each entry of β in turn as the interest parameter β_v.

APPROXIMATE ORTHOGONALISATION

- Seek A^v that makes the vth column of $A^v X$ as orthogonal as possible to its other columns. Write:

$$\begin{aligned} A^v Y &= A^v X\beta + A^v \varepsilon \\ \widetilde{Y}^v &= \widetilde{X}^v \beta + \widetilde{\varepsilon}^v \end{aligned}$$

- If the orthogonalisation is successful, a simple linear regression of $\widetilde{Y}^v$ on $\widetilde{x}_v^v$ (the vth column of $\widetilde{X}^v$) estimates β_v without bias (as in a factorial experiment).

APPROXIMATE ORTHOGONALISATION

- Choose A^v to minimise an observable upper bound on the squared bias plus variance of the resulting estimator.
- Rewriting in terms of $q_v = A^{vT} A^v x_v$ produces a simple unconstrained optimization problem:

$$\operatorname*{argmin}_{q \in \mathbb{R}^n} (q^T x_v)^{-2} q^T (I_n + X_{-v} X_{-v}^T) q.$$

- Such q_v have an exact analytic form, facilitating analysis and comparison.

APPROXIMATE ORTHOGONALISATION

- Exact analytic form for q_v:

$$q_v = a(I_n + X_{-v}X_{-v}^T)^{-1}x_v, \quad a \in \mathbb{R}\backslash\{0\}.$$

- The resulting OLS estimator is $\widetilde{\beta}_v = (q_v^T x_v)^{-1} q_v^T Y$, with bias b_v to be quantified.
- The optimization induces population-level sparsity on the notional Fisher information matrix; $\widetilde{\beta}_v$ exploits this sparsity.

BRIEF COMMENTS

- No penalization, therefore no need to standardize columns of X.
- The procedure is calibrated.
- In a special case, connections to other work can be made.
- A key difference from earlier work is that we induce (do not assume) population-level sparsity on the notional Fisher information matrix. In this sense the approach is closer to Cox and Reid (1987), although it is operationally very different.

	2138	2564	1516	1503	1639	1603	4008	4002	1069	1436	3291	978	3514	1297	1285	3808	1423	1278	403	1290	1303	1312
1	-	-	-	-	-	-	-	-	-	-	-	*	-	*	-	-	-	-	-	*	-	-
2	-	-	-	-	-	-	-	-	*	-	-	-	-	-	*	-	-	*	-	-	-	-
3	-	-	-	-	-	-	-	-	*	-	-	-	-	*	-	-	-	*	-	-	-	-
4	-	-	-	-	-	-	-	-	*	-	-	-	-	-	-	-	*	*	-	-	-	-
5	-	-	-	-	-	-	-	-	*	-	-	-	-	-	*	-	*	-	-	-	-	-
6	-	-	-	-	-	-	-	-	*	-	-	*	-	-	*	-	-	*	-	-	-	-
7	-	-	-	-	-	-	-	-	-	-	-	*	-	*	*	-	-	*	-	-	-	-
8	-	-	-	-	-	-	-	-	-	-	-	*	-	*	*	-	*	-	-	-	-	-
9	-	-	-	*	*	-	-	-	-	-	-	*	-	-	-	*	-	-	-	-	-	-
10	-	-	-	-	-	-	-	-	*	-	-	-	-	*	-	-	-	-	-	-	*	-
11	-	-	-	*	-	-	-	-	*	-	-	-	*	-	-	-	-	-	-	-	-	-
12	-	-	-	*	-	-	-	-	-	*	-	-	*	-	-	-	-	-	-	-	-	-
13	-	-	-	*	-	*	*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
14	-	-	-	-	-	-	-	-	*	-	-	-	-	*	-	-	-	-	-	*	-	-
15	-	-	-	*	*	-	-	-	-	-	-	-	-	-	-	*	-	-	-	-	-	-
16	-	-	-	-	-	-	-	-	*	-	-	-	-	-	-	-	*	-	-	-	-	*
17	-	-	-	-	-	-	-	-	*	-	-	-	-	*	*	-	-	*	-	-	-	-
18	-	-	*	*	-	-	-	-	-	-	-	-	-	-	-	*	-	-	-	-	-	*
19	-	-	-	-	-	-	-	-	*	-	-	*	-	*	*	*	-	-	-	-	-	-
20	-	-	*	-	-	*	-	*	-	-	-	-	-	-	-	-	-	-	-	-	-	-
21	-	-	-	-	-	-	-	-	*	-	-	-	-	*	*	*	-	-	-	-	-	-
22	-	-	-	-	-	-	-	-	*	-	-	-	-	*	-	-	-	-	-	*	*	-
23	-	-	-	-	-	-	-	-	*	-	-	-	-	*	-	-	*	-	-	-	-	*
24	-	-	-	*	*	-	-	-	*	*	-	-	-	-	-	*	-	-	-	-	-	-
25	-	-	*	*	-	-	-	-	-	*	-	-	-	-	-	*	-	-	-	-	-	*
26	-	-	-	-	-	-	-	-	*	-	-	-	-	*	-	-	*	*	-	-	-	-
27	-	-	-	-	*	*	-	*	-	*	-	-	-	-	-	-	-	-	-	-	-	-
28	-	-	-	-	-	-	-	-	*	-	-	*	-	*	*	-	-	*	-	-	-	-
29	-	-	-	*	-	-	-	-	*	-	-	*	-	-	-	-	-	*	-	-	-	*
30	-	-	-	*	-	-	-	-	*	-	-	*	-	-	*	-	-	-	-	-	-	*
31	-	-	-	*	-	-	-	-	*	-	-	*	-	-	-	-	-	-	-	*	-	*
32	-	-	-	-	-	-	-	-	*	-	-	-	-	*	*	-	*	-	-	-	-	*
33	-	-	-	*	*	-	-	-	*	-	*	-	-	-	-	*	-	-	-	-	-	-
34	-	-	*	*	-	-	-	-	-	*	-	-	-	*	-	*	-	-	-	-	-	-

- Usage as an adjunct and refinement to confidence set of models (Cox and Battey, 2017; Battey and Cox, 2018).
- If several or many models fit the data equivalently well, an arbitrary choice among them is fine for prediction but is likely to be misleading for scientific understanding.
- Should aim to report all statistically indistinguishable well-fitting models.

SUMMARY

- For inference on an interest parameter in the presence of a high-dimensional nuisance parameter, the estimation error associated with the nuisance parameter needs to be restrained.
- A common approach is to assume the nuisance parameter array is sparse.
- I have presented examples of recent work with a recurring theme: seek reformulations that induce population-level sparsity in non-standard domains.
- Parameter orthogonalisation (Cox and Reid, 1987) can also be formulated in this way.

The talk was based on:

- Battey, H. S. (2023). Inducement of population sparsity. *Canadian J. Statist. (Festschrift in honour of Nancy Reid)*, to appear.

synthesising:

- Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc., B*, 49, 1–39.
- Battey, H. S. (2017). Eigen structure of a new class of structured covariance and inverse covariance matrices. *Bernoulli*, 23, 3166–3177.
- Battey, H. S. and Cox, D. R. (2020). High-dimensional nuisance parameters: an example from parametric survival analysis. *Information Geometry*, 3, 119–148.
- Battey, H. S. and Reid, N. (2023). On inference in high-dimensional regression. *J. Roy. Statist. Soc., B*, 85, 149–175.
- Battey, H. S., Cox, D. R. and Lee, S. (2023). On partial likelihood and the construction of factorisable transformations. *Information Geometry*, to appear.

Also mentioned but not discussed:

- Rybak, J. and Battey, H. S. (2021). Sparsity induced by covariance transformation: some deterministic and probabilistic results. *Proc. Roy. Soc. Lond. A*, 477.
- Battey, H. S. (2019). On sparsity scales and covariance matrix transformations. *Biometrika*, 106, 605–617.
- Battey, H. S. and Cox, D. R. (2018). Large numbers of explanatory variables: a probabilistic assessment. *Proc. Roy. Soc. London A*, 474.
- Cox, D. R. and Battey, H. S. (2017). Large numbers of explanatory variables, a semi-descriptive analysis. *Proc. Nat. Acad. Sci.*, 114, 8592–8595.