Faculty of Science

Estimation of sparse multivariate diffusions

Alexander Sokol & Niels Richard Hansen
Department of Mathematical Sciences

August 19, 2011

Consider an Ornstein-Uhlenbeck process

$$X_t = x_0 + \int_0^t B(X_s - A)\,\mathrm{d}s + \sigma W_t,$$

where $B, \sigma \in \mathbb{M}(p, p)$ and $A \in \mathbb{R}^p$.

Main question: How do we estimate B in the case where B is sparse, or when B may be decomposed into sparse components?

Applications: Gene regulation networks, et cetera.

Agenda:

1. A nonlinear least squares estimate of B
2. L_1-penalized estimation of B
3. Estimation in the case of symmetric B

A nonlinear least squares estimate of B

The OU process given by $\mathrm{d}X_t = B(X_t - A)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t$ has a pathwise unique solution which is a homegeneous Markov process, and the transition probability $P(h, x, \cdot)$ of transition from state x over time h is a normal distribution with mean and variance

$$\begin{aligned} \xi(h, x) &= A + \exp(hB)(x - A) \\ \Sigma(h) &= \int_0^h \exp(sB)\sigma\sigma^t \exp(sB^t)\,\mathrm{d}s. \end{aligned}$$

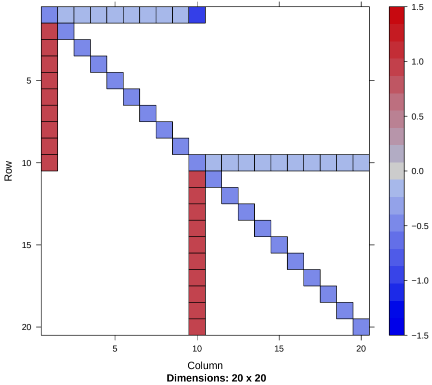
We assume that A and σ are known and take interest in estimating B.

Assume equally spaced observations $t_0, \dots, t_n$, $t_0 = 0$, with step size Δ. We consider the loss function

$$\begin{aligned} L(B) &= \frac{1}{2}\sum_{k=1}^{n} \|X_{t_k} - \xi(\Delta, X_{t_{k-1}})\|_2^2 \\ &= \frac{1}{2}\sum_{k=1}^{n} \|(X_{t_k} - A) - \exp(\Delta B)(X_{t_{k-1}} - A)\|_2^2, \end{aligned}$$

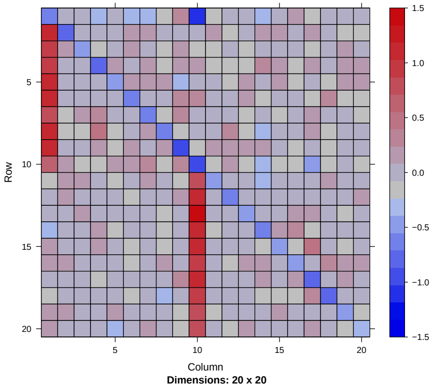
which is loosely related to the minus log likelihood. Even though L is not convex and in general may have nonunique global minima, we may estimate B as a minimizer of L.

We consider an example with $p = 20$, $A = 0$, $x_0 = 0$ and $\sigma = I_p$, and B is the following sparse matrix.



We simulate observations with $T = 100$ and $n = 10000$ and estimate B as a minimizer of $L(B)$.

The result of the numerical minimization:



Conclusions:

The many zero entries of the true parameter are not zero in the estimate.

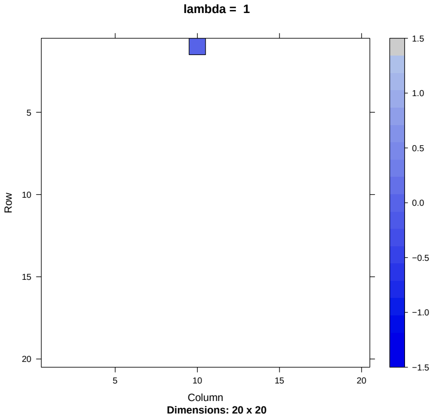
We would like to estimate B in a manner wherein the estimation method actually yields sparse estimates.

L_1-penalized estimation of B

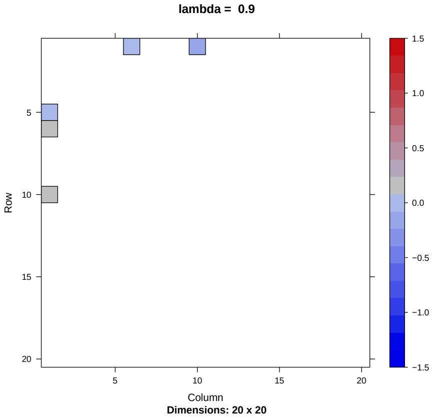
In L_1-penalized estimation, we estimate B as the argument minimum of $L(B) + \lambda\|B\|_1$, where $\|\cdot\|_1$ is the entrywise L_1-norm. The addition of the penalization term has a tendency to yield argument minima which are sparse.

The penalty term implies that we are no longer facing an ordinary nonlinear least squares problem. Nonetheless, linearization of the least squares term combined with the results from L_1-penalized linear regression yields a useful cyclic coordinate algorithm.

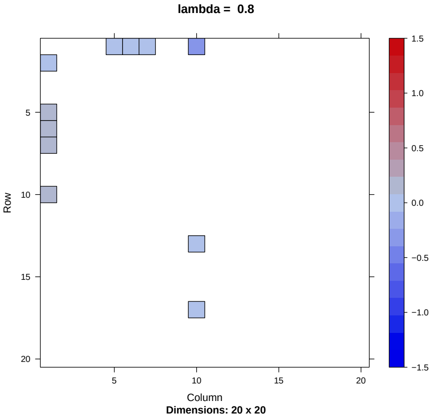
The result of the algorithm:



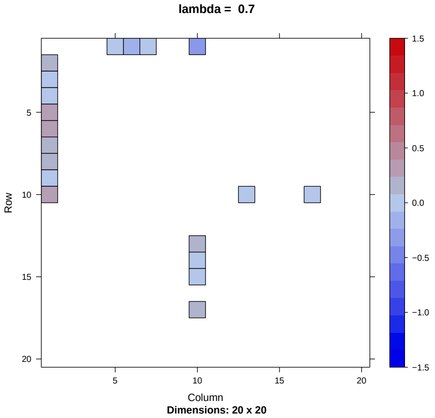
The result of the algorithm:



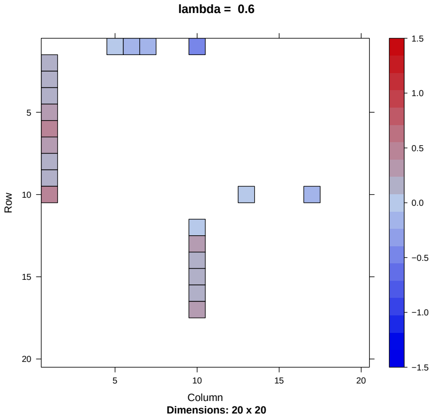
The result of the algorithm:



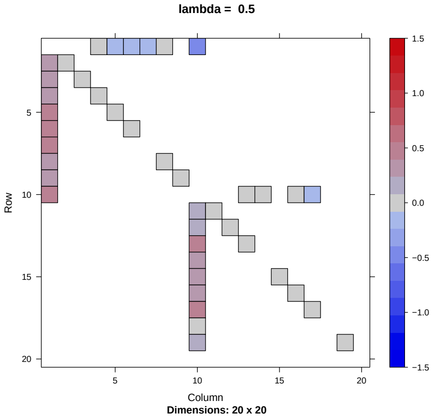
The result of the algorithm:



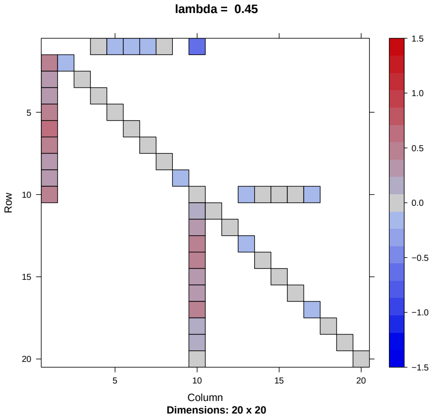
The result of the algorithm:



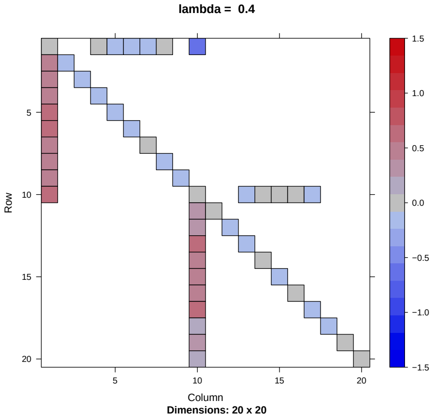
The result of the algorithm:



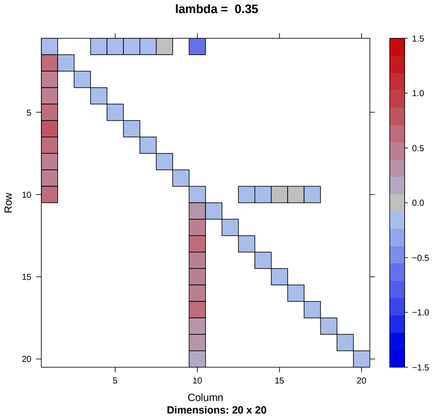
The result of the algorithm:



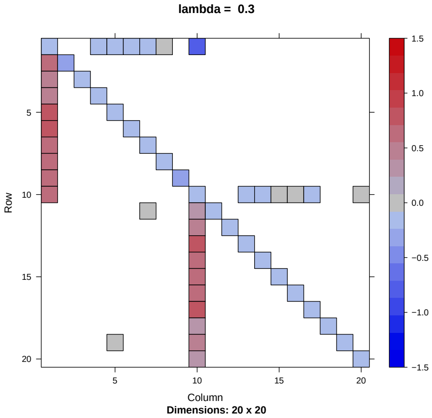
The result of the algorithm:



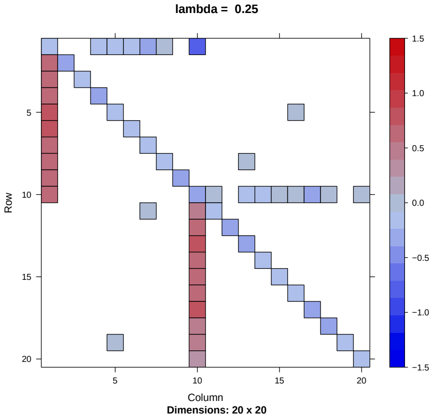
The result of the algorithm:



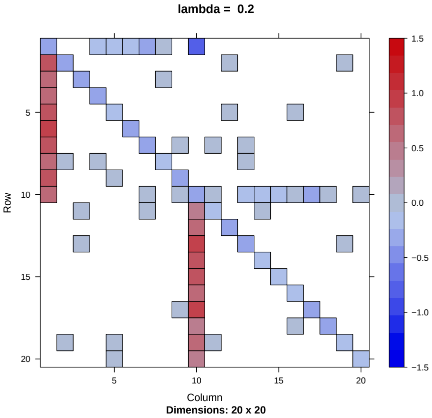
The result of the algorithm:



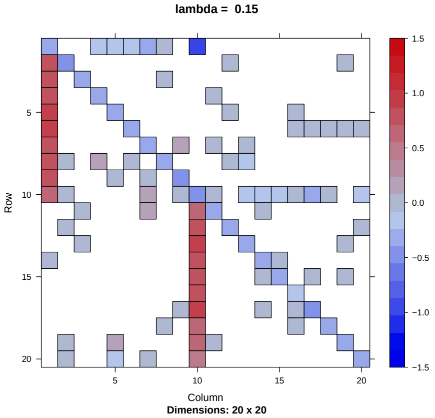
The result of the algorithm:



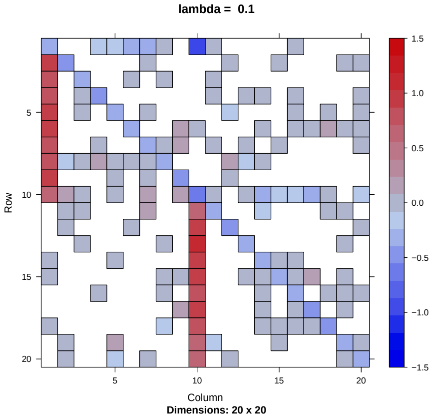
The result of the algorithm:



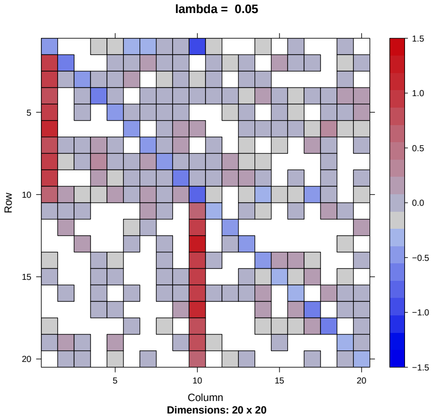
The result of the algorithm:



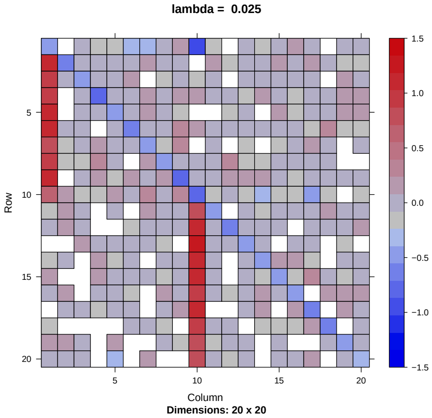
The result of the algorithm:



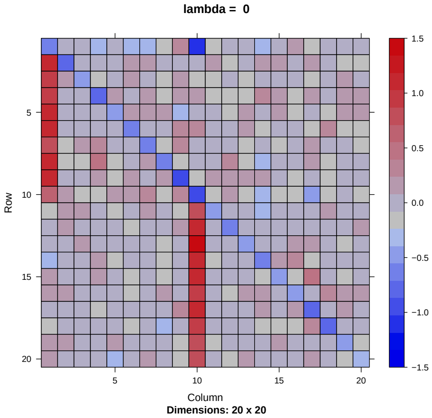
The result of the algorithm:



The result of the algorithm:



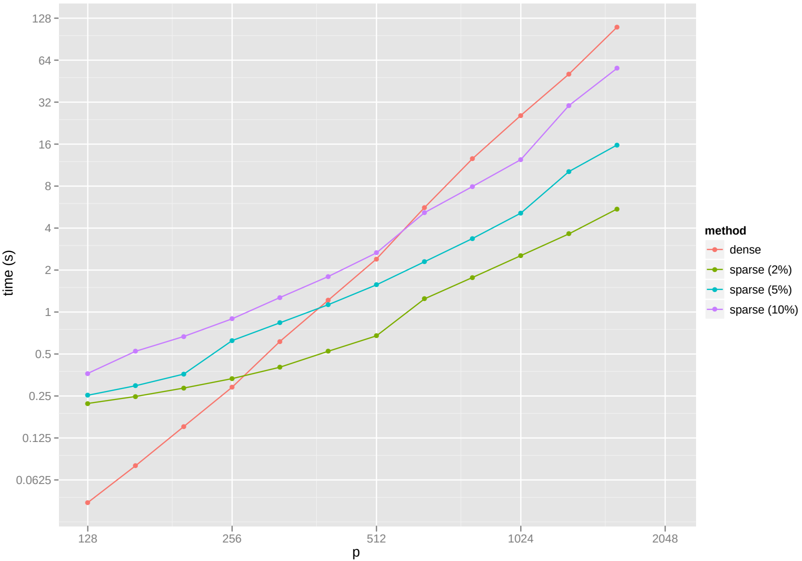
The result of the algorithm:



Conclusions:

1. The L_1-penalized estimates recreate the sparsity of the true parameter at no significant extra numerical cost.
2. In both the ordinary and the L_1-penalized minimization problems, the main computational cost is the calculation of the matrix exponential and its Fréchet derivative.
3. The `R` package `expm` calculates these using Al-Mohy & Higham (2009), not taking advantage of sparsity.
4. A new `R` interface to the Fortran library Expokit speeds up the computation of the matrix exponential for large sparse B-matrices.

Computation time of L:



Estimation in the case of symmetric B

In order to avoid the computational cost of the matrix exponential, we now assume that B is symmetric with $B = PDP^t$, so the loss function becomes

$$L(P, D) = \frac{1}{2} \sum_{k=1}^{n} \|(X_{t_k} - A) - P \exp(\Delta D) P^t (X_{t_{k-1}} - A)\|_2^2.$$

The diagonalization makes $L(P, D)$ simple to calculate, but introduces the problem of optimization of P over the orthogonal group $\mathbb{O}_p$.

While ultimately interested in L_1-penalized estimation, our first step is unpenalized estimation for P and D.

For convenience, consider $L(P, D_e) = \frac{1}{2}\|W - PD_eP^tV\|_2^2$, where $W, V \in \mathbb{M}(p, n)$ and $D_e = \exp(\Delta D)$. We want to minimize $L(P, D_e)$ over $P \in \mathbb{O}_p$ and $D_e \in \mathrm{diag}_p$.

For fixed P, $L(P, D_e)$ has a unique minimum over D_e given by, with d_e denoting the vector of diagonal entries,

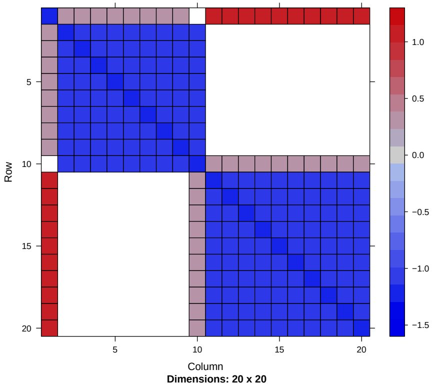
$$(d_e)_k = \frac{\langle (P^tV)_{k\cdot}, (P^tW)_{k\cdot}\rangle}{\|(P^tV)_{k\cdot}\|_2^2}$$

The main problem is therefore minimization over $P \in \mathbb{O}_p$.

A gradient descent algorithm for minimizing $L(P, D_e)$ over $P \in \mathbb{O}_p$:

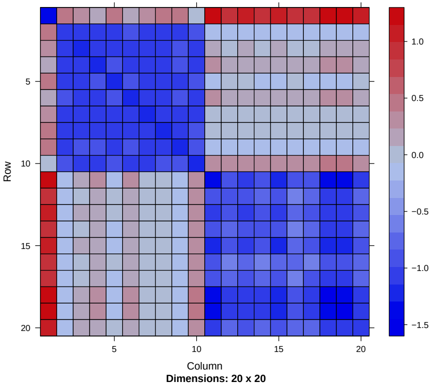
1. Pick initial guess $P^* \in \mathbb{O}_p$.
2. Calculate the gradient $\mathbb{D}_P L(P^*)$ as if the domain were $\mathbb{M}(p, p)$ and not $\mathbb{O}_p$. Calculate projection $\mathbb{D}_P^{\mathbb{O}_p} L(P^*)$ onto the tangent space of P^*. Pick step size by "inexact" minimization of $L(P^* - h\mathbb{D}_P^{\mathbb{O}_p} L(P^*))$.
3. Use singular value decomposition to update P^* as the projection of $P^* - h\mathbb{D}_P^{\mathbb{O}_p} L(P^*)$ onto $\mathbb{O}_p$.
4. Repeat step 2 and step 3 until convergence.

We consider a symmetric example with $p = 20$, $A = 0$, $x_0 = 0$ and $\sigma = I_p$, and B is the following matrix.



We simulate observations with $T = 100$ and $n = 10000$.

Estimate from gradient descent algorithm, 1000 iterations, given true D_e:



Current status:

1. Parametrization by (P, D) avoids expensive matrix exponentiation.
2. Problems with estimation without fixed eigenvalues.
3. There are many aspects of estimation over $\mathbb{O}_p$ which are open to questioning, such as nonconnectivity of $\mathbb{O}_p$, optimization along geodesics, step size selection etc.
4. Practical solutions under L_1-penalization still lacking.

Thank you