1A1-2

大規模次元システムのモデリング手法の提案とその性能評価

A study on high dimensional modeling

ベトフォン グエン[*1] 鷲尾 隆[*1]
Nguyen Viet Phuong Takashi Washio

[*1]大阪大学産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

This paper assesses a high dimensional modeling approach. The performance of the proposed approach is evaluated by articially generated data.

1. はじめに

近年，IT 技術の発展により，計算機のみならず様々な情報機器，家電，自動車，社会インフラなどがネットワークで結ばれつつある．このネットワーク化したユビキタス社会において，様々な目的で個々の対象にセンサーが付けられデータ収集が行われている．その際，通信コスト節約の理由で多数のシステムではセンサーが定期的ではなく駆動的に動作する．即ち，事象が起こった時だけその事象に対応するセンサーが反応し発信する．例えば，特定場所に観測センサーが設置されている交通システムにおいては，車の流量がある数以上であると観測した時のみ，センサーが警告発信する場合が多い．そのため，センサーをはじめシステムに設置されるデータ収集装置群全体からバイナリ不等間隔なトランザクション時系列データが創出される場合が多い．対象システムについて円滑な制御，管理，異常診断などをするためには，このようなデータが表す対象のダイナミクスモデルの構築が必要となる．しかし，このようなモデル化手法の研究は世界的にも十分には行われていない．

対象システムのダイナミクスを遷移遷移により表す古典的な隠れ状態マルコフモデルやマルコフ連鎖モデルが良く知られている．また，近似を導入して状態空間および状態遷移の組み合わせ数を削減することにより高次マルコフ連鎖を効率的に表すモデルである Mixture Transition Distribution (MTD) [Berchtold 2002] や Variable Length Markov Chain (VMLC) [Bejerano 2001] が提案されている．しかし，これらのモデルでは与えられた個々の過去状態から次の状態に遷移する確率分布が提供されるのみであり，対象システムが時間の共にどのように変化するのかに関する特性は表されない．例えば，上記の交通システムの例にマルコフ連鎖などのモデルを適用する場合，流量状態の遷移確率が分かるが，流量を変化させる重要な交通のメカニズムがモデル化されない．一方，これに対して，対象システムのメカニズムを明確にモデル化可能な手法としてカルマンフィルター [Kalman 1960] とその拡張 [Julier 1997] が考えられるが，時系列連続観測データを扱えるのみであり前述のような観測バイナリトランザクション時系列データのみから対象ダイナミクスをモデル化することは困難である．

そこで，本研究では，上記の従来手法の問題を克服し，観測バイナリ時系列データから対象システムのダイナミクスを表すモデルを得るモデリング手法を提案する．

連絡先:
氏名：Nguyen Viet Phuong
所属：大阪大学産業科学研究所
住所：大阪府茨木市美穂ヶ丘 8-1
電子メールアドレス：vphuong@ar.sanken.osaka-u.ac.jp

2. 提案モデルの定式化

2.1 データ形式および状態の定義

N 個のセンサーが設置される対象システムを想定する．T ステップに亘る観測バイナリトランザクション時系列データは

$$\mathbf{D} = \mathbf{Q_1}\mathbf{Q_2}...\mathbf{Q_T}$$

と記述される．但し，$\mathbf{Q_t}(t = 1, 2, ..., T)$ は時刻 t で観測された N 次元バイナリベクトル $\mathbf{Q_t} = (q_t^i)_{i=1,2,..,N}$ である．時刻 t においてセンサー i が発信すれば $q_t^i = 1$ であり，そうでなければ $q_t^i = 0$ である．一方，時刻 t の対象システム状態を，状態ベクトル $\mathbf{X_t} = (x_t^i)_{i=1,2,..,N}$ で定義する．各 x_t^i は $(-\infty < x_t^i < \infty)$ であり，センサー i の発信可能性を決める仮想の測度である．x_t^i が大きければセンサー i の発信可能性が高く，逆に小さければ低い．

2.2 モデルの定式化

提案モデルは次の 2 つ方程式からなる．

$$\mathbf{X_t} = \mathbf{A} \times \mathbf{X_{t-1}} \quad (1)$$

$$\mathbf{Q_t} = H(Sm(\mathbf{X_t})) \quad (2)$$

式 (1) はシステムダイナミクスを表す状態間遷移過程である．本稿では線形ダイナミクスのモデルを対象とするので，$\mathbf{A} = (a_{ij})_{N\times N}$ は $N \times N$ 行列であり，状態遷移行列と呼ばれる．式 (2) はシステム状態 $\mathbf{X_t}$ からセンサーの発信記録バイナリベクトル $\mathbf{Q_t}$ への写像を表す観測過程である．ここで，$Sm()$ はシグモイド関数で，

$$Sm(x) = \frac{1}{1 + \exp(-x)}$$

である．但し $x \in \Re, Sm(x) \in [0, 1]$ である．H は $[0, 1]$ から $\{0, 1\}$ への写像である．$H(p)$ は $[0, 1]$ で一様乱数 h を生成し，$0 \le h \le p$ の場合 $H(p) = 1$, $p < h \le 1$ の場合 $H(p) = 0$ を出力する．シグモイド関数はセンサーが発信する可能性を表す仮想的な変数 $x \in \Re$ を確率変数 $p \in [0, 1]$ に変換する．H のオペレターにより，発信記録バイナリ値 (0,1) を決める．

3. モデリング手法

以上より，提案モデルは状態遷移行列 $\mathbf{A} = (a_{ij})_{N\times N}$ で構成される．行列 $\mathbf{A}$ が分かれば，システムのダイナミクスの特性が明確になる．すなわち，観測バイナリトランザクション時系列データのモデリングは，$\mathbf{Q_t} = (q_t^i)_{i=1,2,..,N}(t = 1, 2, ..., T)$ からの行列 A を推定することである．しかし，データから直接に遷移行列 A を推定することが難しいので，状態ベクトル系列 $\mathbf{X_t}$ の推定と並行して反復計算によりもとめる．

3.1 遷移行列が既知である時の状態ベクトル系列の推定

システム状態遷移行列 A が与えられた時，時刻 t において観測バイナリデータ $\mathbf{Q_{t-1}}, \mathbf{Q_t}$ 下でシステム状態ベクトル $\mathbf{X_t}$ は事後確率 $P(\mathbf{X_t}|\mathbf{Q_{t-1}}, \mathbf{Q_t})$ を最大化することにより決定される．確率計算により，$P(\mathbf{X_t}|\mathbf{Q_{t-1}}, \mathbf{Q_t})$ は以下となる．

$$P(\mathbf{X}_t|\mathbf{Q}_{t-1}, \mathbf{Q}_t) = \frac{P(\bar{\mathbf{X}}_{t-1})R_{t-1}(\bar{\mathbf{X}}_{t-1})R_t(\mathbf{X}_t)}{P(\mathbf{Q}_t|\mathbf{Q}_{t-1})P(\mathbf{Q}_{t-1})} \tag{3}$$

但し，

$$R_t(\mathbf{X}_t) = \prod_i^N \left(q_t^i Sm(x_t^i) + (1 - q_t^i)\{1 - Sm(x_t^i)\}\right),$$
$$R_{t-1}(\bar{\mathbf{X}}_{t-1}) = \prod_i^N \left(q_{t-1}^i Sm(\bar{x}_{t-1}^i) + (1 - q_{t-1}^i)\{1 - Sm(\bar{x}_{t-1}^i)\}\right),$$
$$\bar{x}_{t-1}^i = \sum_j^N \left[a'_{ij}x_t^j\right], (a'_{ij})_{N\times N} = \mathbf{A}^{-1}.$$

定理１: 時刻 t において，状態ベクトル $\mathbf{X_t}$ の各要素 x_t^i が独立で $(-\infty, \infty)$ の値を取り，x_t^i の p.d.f が $P(x_t^i) = f(x_t^i)$ である場合，式 (3) の $P(\mathbf{X_t}|\mathbf{Q_{t-1}}, \mathbf{Q_t})$ が最大となる $\mathbf{X_t}$ は以下の非線形連立方程式の解である．

$$\begin{aligned} &q_t^k - Sm(x_t^k) + \sum_{i=1}^N a'_{ik}\left(q_{t-1}^i - Sm[\sum_{j=1}^N a'_{ij}x_t^j]\right) \\ &+ \sum_{i=1}^N \frac{a'_{ik}}{f(\bar{x}_{t-1}^i)} f'(\bar{x}_{t-1}^i) = 0, (k = 1, 2, ..., N) \end{aligned} \tag{4}$$

（証明略）

定理２: $f(x_t^i)$ が２次微分できる場合，式 (4) の解が存在し，常に Newton 法より収束計算可能である．（証明略）

定理 1,2 より，状態遷移行列 $\mathbf{A}$ および観測バイナリデータ $\mathbf{Q_t}$ から状態ベクトル系列 $\mathbf{X_t}$ を導出できる．

3.2 状態遷移行列の導出

状態ベクトル系列 $\mathbf{X_t}(t = 1, 2, .., N)$ が与えれた時，一義的には予測誤差

$$L = \sum_{t=1}^T \|\mathbf{X}_t - (\mathbf{A} \times \mathbf{X}_{t-1})\|_2^2$$

を最小化することにより $\mathbf{A}$ を推定可能である $\|\bullet\|_2^2$ は２次ベクトルのノルムである．しかし，一般に状態間の影響は局所的であることが多く，状態遷移行列 $\mathbf{A}$ はスパースであるという現実的仮定を行うことで，L_1-正則化法 [Meinshausen 2006] を用い行列 $\mathbf{A}$ を導出する．L1 は

$$L_1 = \sum_{t=1}^T \|\mathbf{X}_t - (\mathbf{A} \times \mathbf{X}_{t-1})\|_2^2 + \gamma \|\mathbf{A}\| \tag{5}$$

である．$\|\mathbf{A}\|$ は行列 $\mathbf{A}$ のノルムであり，$\|\mathbf{A}\| = \sum_{i=1}^N \sum_{j=1}^N |a_{ij}|$ で定義する．$\gamma > 0$ は正則化パラメータである．

定理３: 与えられる状態ベクトル系列 $\mathbf{X_t}$ が正確であれば，時系列の長さ T が大きくなるほど，L_1 を最小化する推定遷移行列 $\hat{\mathbf{A}}$ は真の遷移行列に近づく．但し，γ の値を大きくすれば $\hat{\mathbf{A}}$ はよりスパースとなる．（証明略）

この定理により，状態ベクトル系列から真に近い遷移行列 $\mathbf{A}$ を導出することが可能と考えられる．但し，L_1 正則化の解をもとめるためには条件付の最小二乗法問題 [Lawson 1977] に帰着し既存のアルゴリズムを使用する．詳細アルゴリズムは [Tibshirani 1996] で提案されている．

3.3 観測データからの遷移行列 A の構成

節 3.2, 3.3 の結果に基づいて，観測バイナリ時系列データからの以下のようなモデリング手順を提案する．

ステップ１ 初期の遷移行列 $\mathbf{A}$ を仮定する．

ステップ２ 観測バイナリ時系列データ $\mathbf{Q_t} = (q_t^i)_{i=1,2,..,N}$ から定理２により状態ベクトル系列を推定する．

ステップ３ 推定した状態ベクトルから式 (5) の L_1 正則化条件を満たす行列を導出し，遷移行列を更新する．

ステップ２とステップ３を収束するまで繰り返し行えばモデルのダイナミクスを表す遷移行列 $\mathbf{A}$ が得られる．

4. 性能検証

4.1 検証用人工バイナリ時系列データの生成

提案するモデリング手法の性能検証用のバイナリ時系列データをシミュレーションにより人工的に作成した．システムの変数個数（センサーの数）を N とする．シミュレーションプログラムは，人工システムの実数遷移行列 $\mathbf{SimM}_{N\times N}$ を自動的に構成する．ここで，様々な特性のシステムダイナミクスに対するモデリング評価を行うため，遷移行列 $\mathbf{SimM}$ は既知の固有値集合に基づいて作られる．固有値集合 $\{\lambda_1, \lambda_2, .., \lambda_N\}$ が与えれば，ランダムに $N \times N$ の正則行列 $\mathbf{P}$ を生成し，

$$\mathbf{SimM} = \mathbf{P\Lambda P}^{-1}$$

により遷移行列 $\mathbf{SimM}$ を計算する．但し，$\mathbf{\Lambda} = diag(\lambda_i), (i = 1, 2, ..., N)$ である．また，固有値が複素数となることも許す．乱数で生成する初期状態ベクトルから初めて式 (1) のシステム過程により過去の状態から次の状態を T ステップで繰り返し，人工の状態ベクトル系列 $\mathbf{SimX_t}(t = 1, 2, ..., T)$ を計算する．次に，各時間ステップでは，式 (2) の観測過程を用いて状態ベクトルから人工バイナリ時系列データ $\mathbf{SimD}$ をランダムに生成する．

4.2 モデリング結果の評価法

シミュレーションで生成した人工バイナリトランザクション時系列に提案モデリング手法を適用し，状態遷移行列 $\mathbf{A}$ を推定する．そして，この状態遷移行列 $\mathbf{A}$ を元のシミュレーション遷移行列 $\mathbf{SimM}$ と比較して遷移行列の再現性を評価する．しかし，対象システムの特性に注目するため，行列の各要素ではなく行列が持つ固有値の誤差により検証を行う．ここで，いくつかの誤差指標を導入する．

固有値の誤差 E^λ: 固有値 λ_x に関する推定値 λ_y の誤差を一般的に

$$E^\lambda(\lambda_x, \lambda_y) = |Re(\lambda_x) - Re(\lambda_y)| + |Im(\lambda_x) - Im(\lambda_y)| \tag{6}$$

で定義する．Re は実部であり Im は虚部を表す．

固有値集合の誤差 E: 二つの固有値の集合 $H = \{\lambda_I | I = 1, 2, ..., N\}$，$H' = \{\lambda'_J | J = 1, 2, .., N\}$ が与えられているとする．今，H の固有値集合から H' の固有値集合への全単写像 $F : H \to H'$ の誤差を

$$\sum_{\forall \lambda_I \in H, \lambda'_J = F(\lambda_I)} E^{\lambda}(\lambda_I, \lambda'_J)$$

と定義し，更に上記が最小となる条件を満たす全単写像を F_{min} とする．その時，平均誤差を

$$\overline{\mathrm{E(H, H')}} = \frac{1}{N} \sum_{\forall \lambda_I \in H, \lambda'_J = F_{min}(\lambda_I)} E^{\lambda}(\lambda_I, \lambda'_J) \qquad (7)$$

と定義する．

シミュレーションの遷移行列 **SimM** が持つ固有値の集合を $H = \{\lambda_I^{SimM} | I = 1, 2, ..., N\}$ とし，モデリングで得られた **A** の固有値集合を $H' = \{\lambda_J^A | J = 1, 2, .., N\}$ とすれば，誤差は式 (7) となる．しかしながら，実数と複素数固有値の推定能力を区別しながら評価するため，**SimM** の固有値集合を実数固有値集合 $\lambda^{SimM,r} = \{\lambda_1^{SimM,r}, \lambda_2^{SimM,r}, ..., \lambda_{Nr}^{SimM,r}\}$ および複素数固有値集合 $\lambda^{SimM,c} = \{\lambda_1^{SimM,c}, \lambda_2^{SimM,c}, ..., \lambda_{Nc}^{SimM,c}\}$ に分ける（$Nr + Nc = N$）．実数の固有値，複素数固有の推定平均誤差はそれぞれ，

$$\overline{E_{real}} = \frac{1}{Nr} \sum_{i=1}^{Nr} E^{\lambda}(\lambda_i^{SimM,r}, F_{min}(\lambda_i^{SimM,r}))$$
$$\overline{E_{complex}} = \frac{1}{Nc} \sum_{i=1}^{Nc} E^{\lambda}(\lambda_i^{SimM,c}, F_{min}(\lambda_i^{SimM,c}))$$

とする．その時，式 (7) は $\overline{E} = \frac{Nr}{N}.\overline{E_{real}} + \frac{Nc}{N}.\overline{E_{complex}}$ となる．

4.3 実験評価結果

様々なシステムの特性に対する同定能力の検討をするために，変数の数 N を 3，4 とし時間ステップ数 T を 1000 として，遷移行列の固有値集合を表 1 のように設定して 16 通りの人工バイナリ時系列データを生成した．但し，$N = 3$ の場合，$\lambda_1, \lambda_2, \lambda_3$ を用い，$N = 4$ の場合は λ_4 も用いる．ここでは，対象が安定なシステムである場合のみを考察するため，全ての固有値の絶対値を 1 以下としている．$N = 3$ の場合，L_1 正則化のパラメータ γ を 0.2 とし，$N = 4$ の場合，$\gamma = 0.125$ に設定しモデリング手法を適用した．得られたモデル遷移行列の固有値誤差を計算した結果を図 1，2 に示す．シミュレーション遷移行列 **SimM** が絶対値 1 となる複素数固有値を持つ場合 (セット#1,#2,#3,#4,#5,#6)，複素固有値の平均誤差 $\overline{E_{complex}}$ は小さいが，実数固有値の平均誤差 $\overline{E_{real}}$ は相当大きい．絶対値 1 の複素固有値についてはシミュレーションシステムの状態ベクトルが周期的に減衰せず振動する．従って，その特性がバイナリデータの全体に反映されるので，ほぼ正確にモデル推定可能である．一方，実数固有値は 1 より小さいので，シミュレーションの状態遷移では短いステップでそれを反映する挙動が消える．そのため，モデル推定は難しい．#7 と #8 の固有値セットに対する誤差もこの議論を補強する．#7 の場合，複素固有値はなく振動しないが，実数固有値 1 によりシステム状態ベクトルが長い時間経てば遷移行列の固有値 1 に対応する固有ベクトルに収束しバイナリデータはほぼ一定となる．逆に，#8 の場合は，全ての固有値の絶対値が 1 より小さいので，状態ベクトルが早く 0 になり，観測過程はほぼランダムになりモデル化が困難である．この結果により，提案したモデリング手法では，複素数固有値に対応する振動などを表すシステムの主要ダイナミクスの特性を正しく同定することは可能であるが，小さな実数固有値に対応する挙動の推定は難しいと考えられる．

表 1: シミュレーション遷移行列の固有値集合

セット	λ_1	λ_2	λ_3	λ_4
#1	0.198+0.980i	0.198-0.980	0.823	0.413
#2	0.405+0.915i	0.405-0.915i	0.823	0.413
#3	0.598+0.802i	0.598-0.820i	0.823	0.413
#4	0.790+0.612i	0.790-0.612i	0.823	0.413
#5	0.921+0.391i	0.921-0.391i	0.823	0.413
#6	0.982+0.191i	0.982-0.191i	0.823	0.413
#7	1	0.445	0.072	0.413
#8	0.404+0.752i	0.404-0.752i	0.823	0.413

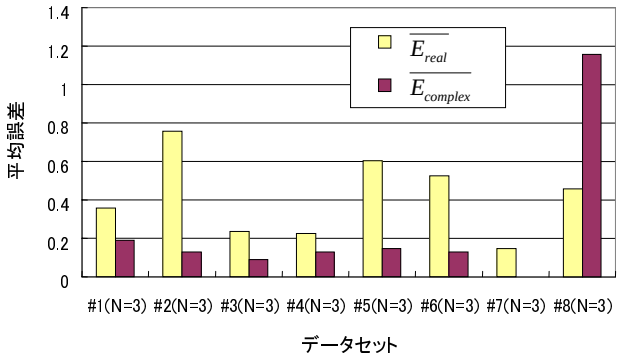

図 1: 各システム特性に対する平均誤差 ($N = 3$)

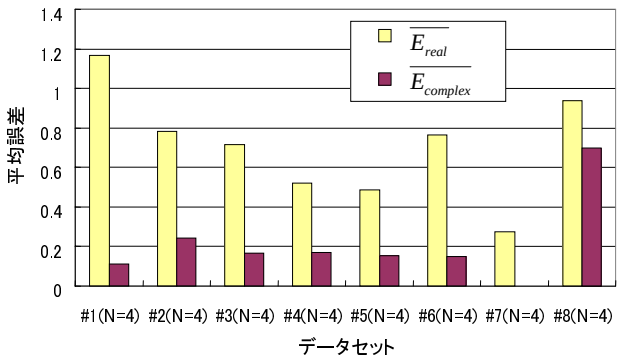

図 2: 各システム特性に対する平均誤差 ($N = 4$)

次に，本モデリング手法によって同定された状態遷移行列の固有値が，シミュレーションで用いた真の行列の固有値からどのように偏移するかを考察する．図 3, 4 はシミュレーション遷移行列および推定遷移行列の複素固有値のプロットである．四角印 (シミュレーション固有値) と三角印（推定した固有値）の間のバーは誤差を表す．図 3, 4 より全体として誤差は小さいが，特に複素固有値の偏角が大きいほど誤差が小さくなる傾向が見られる．偏角が大きい固有値は状態の振動周期が短い．

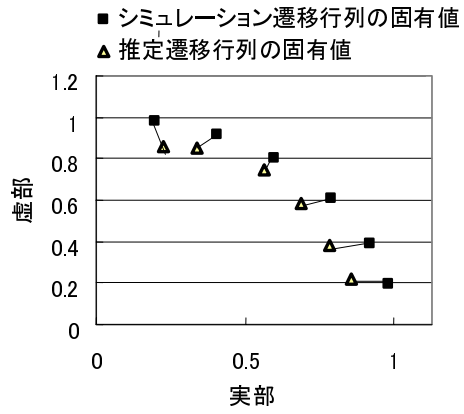

図 3: システム特性を表す複素固有値の推定精度 ($N = 3$)

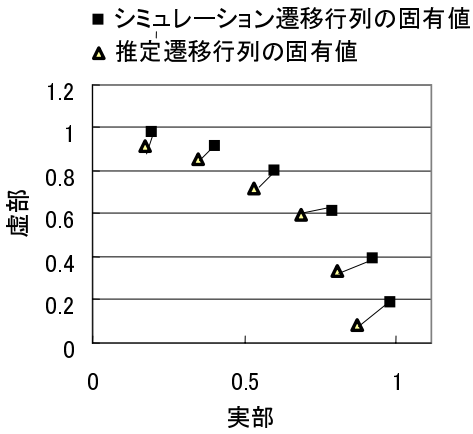

図 4: システム特性を表す複素固有値の推定精度 ($N = 4$)

これは一時間区間中により多くの振動変動が現れ，遷移特性が明確であるため，そのダイナミクスを獲得しやすいからと考えられる．

図 5 は，#2 のデータについて，データの時間ステップ数 T に対して同定精度がどう変化するを示している．これより，事例数が増えると複素固有値に関する平均誤差 $\overline{E_{complex}}$ が減少し，反対に 1 より小さい実数の固有値に関する誤差 $\overline{E_{real}}$ が大きくなる傾向が見られる．定理 3 により T が大きく状態ベクトルが正しければ L_1 正則化を用いて真の遷移行列により近い遷移行列を推定できる．その理由で $\overline{E_{complex}}$ が下がる．しかしながら，単調減衰する挙動を示す実数固有値に関しては，挙動の定常性が成立せず精度は改善されない．

5. まとめ

本稿では線形ダイナミクスシステムからランダム過程により観測されるバイナリトランザクション時系列データの確率的状態遷移モデルとそのモデリング手法を提案した．このアプローチの重要な点はバイナリデータのみから対象システムにおいて働いているダイナミクスモデルを発掘できることである．今後は実用化に向けて，提案手法をより高度化し，大きい変数個数 N を持つ大規模システムに適用可能とする予定である．

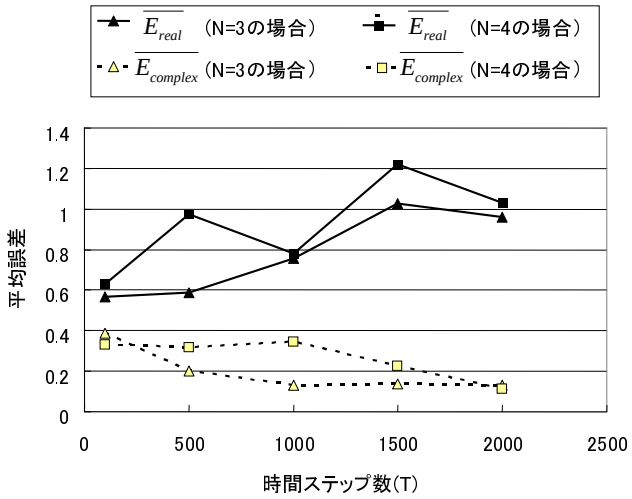

図 5: データの時間ステップ数 T に対する平均誤差

参考文献

[Bejerano 2001] Bejerano, G and Yona, G: Variations on probabilistic suffix trees - a new tool for statistical modeling and prediction of protein families. Bioinformatics, 17(1). (2001)

[Berchtold 2002] Berchtold, A and Raftery, A. E: The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. Statistical Science, 17(3). (2002)

[Julier 1997] Julier, Simon, J and Jeffery, K. U: A New Extension of the Kalman Filter to nonlinear Systems. In the Proceedings of AeroSense: The 11th International Symposium on Aerospace/Defense Sesing, Simulation and Controls, Multi Sensor Fusion, Tracking and Resource Management II, SPIE, 1997.

[Kalman 1960] Kalman, R. E.: A New Approach to Linear Filtering and Predition Problems. Transactions of the ASME - Journal of Basic Engineering, 82. (1960)

[Lawson 1977] Lawson, C and Hansen, R: Solving least squares problems, Journal of the American Statistical Association, 72(360). (1977)

[Meinshausen 2006] Meinshausen, N and Buhlmann, P: High-Dimensional Graphs and Variable Selection with the LASSO, The Annuals of Statistic, 34(3). (2006)

[Tibshirani 1996] Tibshirani, R: Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B., 58(1). (1996)