Factoring Gaussian Precision Matrices for Linear Dynamic Models

Joe Frankel and Simon King

Centre for Speech Technology Research
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
Tel: +44 131 651 1769
Fax: +44 131 650 4587

Abstract

The linear dynamic model (LDM), also known as the Kalman filter model, has been the subject of research in the engineering, control, and more recently, machine learning and speech technology communities. The Gaussian noise processes are usually assumed to have diagonal, or occasionally full, covariance matrices. A number of recent papers have considered modelling the precision rather than covariance matrix of a Gaussian distribution, and this work applies such ideas to the LDM. A Gaussian precision matrix P can be factored into the form $P = U^T SU$ where U is a transform and S a diagonal matrix. By varying the form of U, the covariance can be specified as being diagonal or full, or used to model a given set of spatial dependencies. Furthermore, the transform and scaling components can be shared between models, allowing richer distributions with only marginally more parameters than required to specify diagonal covariances.

The method described in this paper allows the construction of models with an appropriate number of parameters for the amount of available training data. We provide illustrative experimental results on synthetic and real speech data in which models with factored precision matrices and automatically-selected numbers of parameters are as good as or better than models with diagonal covariances on small data sets and as good as models with full covariance matrices on larger data sets.

Key words:
Linear dynamic model, error distribution, precision matrix

Email address: joe@cstr.ed.ac.uk (Joe Frankel and Simon King).
URL: http://www.cstr.ed.ac.uk/∼joe/ (Joe Frankel and Simon King).

1 Introduction

The method presented here was developed on an automatic speech recognition (ASR) task, but is applicable to any task using linear dynamical models (LDMs). The method's key property is that it allows the construction of models with a number of parameters in between those of models using diagonal and full covariance matrices; the number of parameters can be automatically determined. In situations without sufficient data to estimate full covariance matrices, the method allows modelling of just some of the spatial correlations, which can be significantly better than using diagonal covariances.

ASR systems frequently employ mixture Gaussian distributions to model the acoustic features associated with each hidden Markov model (HMM) state (Young, 1995; Gold and Morgan, 1999). Extraction of speech features such as Mel-frequency cepstral coefficients (MFCCs) includes steps which reduce (though do not entirely remove) the correlations between dimensions of the feature vectors (Macho et al., 1999). To reduce parameterization and computation time, Gaussian components are frequently estimated to have diagonal rather than full covariance matrices.

The first plot of Figure 1 shows a set of 2-dimensional spatially correlated data, sampled from a multivariate Gaussian distribution. The second plot shows the same data, but reflected in the y-axis. The parameters of diagonal and full covariance Gaussian distributions were estimated from the data, and in each plot single standard deviations from the mean of their output distributions are shown by the dashed and full ellipses respectively. The full covariance models are able to rotate the principal axes of the distribution and therefore give a more informative description of the data. Furthermore, with one set of data a reflection of the other, identical diagonal covariance distributions are estimated for each, and classification of unseen data based on those models is impossible. This illustrates some advantages of modelling the dependencies in spatially correlated data.

1.1 Factoring Gaussian precision matrices

The probability density function (pdf) of a p-dimensional Gaussian-distributed random variable $\mathbf{y} \sim N(\boldsymbol{\mu}, \Sigma)$ is defined by:

$$f(\mathbf{y}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\} \tag{1}$$

The covariance matrix Σ, and hence the precision $P = \Sigma^{-1}$ are symmetric positive definite, a property which makes a factorization of the form $P =$

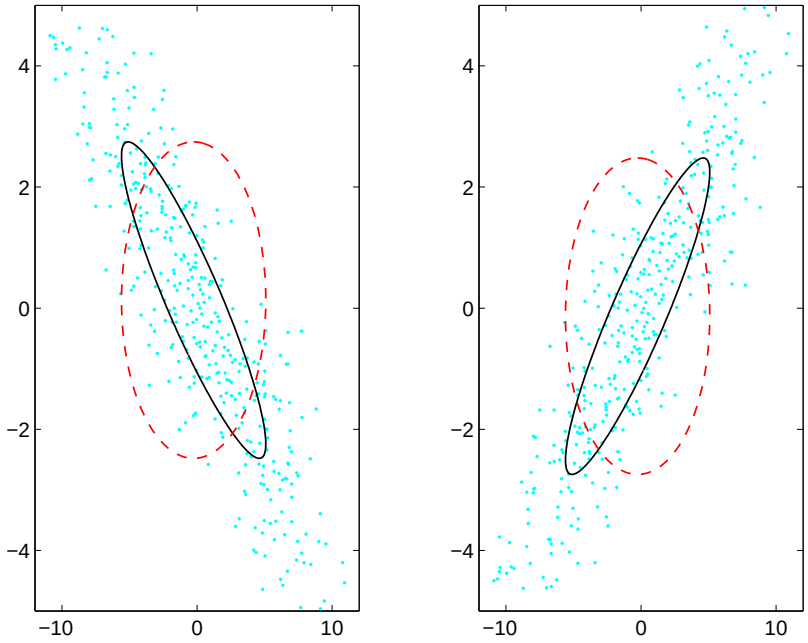

Fig. 1. Single standard deviation ellipses for diagonal and full covariance multivariate Gaussian models estimated on spatially correlated data. The data in the second plot are a reflection of those in the first.

$U^T SU$ possible, where S is a diagonal matrix, and U is a transform. The distribution described in Equation 1 can then be written as:

$$f(\mathbf{y}|\boldsymbol{\mu}, U, S) = \frac{|U^T S U|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T U^T S U(\mathbf{y}-\boldsymbol{\mu})\right\} \quad (2)$$

$$= \frac{|U||S|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}\left(U(\mathbf{y}-\boldsymbol{\mu})\right)^T S \left(U(\mathbf{y}-\boldsymbol{\mu})\right)\right\} \quad (3)$$

Factoring the precision matrix into a combination of transform and diagonal components provides the possibility to model none, all, or a subset of the spatial dependencies present in a set of features. Furthermore, the transforms can be shared between Gaussian components, giving richer models with only a modest increase over the number of free parameters required to specify a diagonal covariance.

Gales (1999) first applied such a factorization to Gaussian precision matrices for ASR. This was a development of state-specific rotations as proposed by Ljolje (1994), in which single full Gaussian covariances were estimated for each state and used to derive decorrelating transforms. The transforms were then fixed and applied to the features prior to estimating diagonal-component Gaussian mixture models (GMMs) for each state.

Under Gales' formulation, direct optimization of U's partial derivative,

$$\frac{\partial \log f}{\partial U} = U^{-T} - SU(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^T \quad (4)$$

where U^{-T} denotes the transpose of U^{-1}, is complex. Instead, an iterative scheme is presented which, for U either full or block-diagonal, guarantees to increase the model likelihood on the training data.

Olsen and Gopinath (2004) generalizes this approach by generating the precision matrix from a set of D basis elements such that $P = \sum_{k=1}^{D} \lambda_k \mathbf{a}_k \mathbf{a}_k^T$ for scalar λ_k and p-dimensional vector $\mathbf{a}_k$. By altering the number and makeup of the basis elements, the covariance can be varied from diagonal to full.

Bilmes (2000) simplifies parameter estimation of such models by constraining the form of U to be unit upper-diagonal. Along with ensuring that the factorization of $P = \Sigma^{-1}$ is unique, under such a model $|U| = 1$, so that the specification of the distribution reduces to:

$$f(\mathbf{y}|\boldsymbol{\mu}, U, S) = \frac{|S|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}\left(U(\mathbf{y}-\boldsymbol{\mu})\right)^T S \left(U(\mathbf{y}-\boldsymbol{\mu})\right)\right\} \quad (5)$$

and the partial derivative with respect to U is now linear in U. With I_p denoting a p-dimensioned identity matrix, setting $U = B + I_p$, so that B contains the off-diagonal elements of U and zeros along the diagonal, and $\boldsymbol{\mu}' = U\boldsymbol{\mu}$, Equation 5 can be rewritten as:

$$f(\mathbf{y}|\boldsymbol{\mu}, B, S) = \frac{|S|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{y} + B\mathbf{y} - \boldsymbol{\mu}'\right)^T S \left(\mathbf{y} + B\mathbf{y} - \boldsymbol{\mu}'\right)\right\} \quad (6)$$

$$= \frac{|S|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}\sum_{l=1}^{p} s_{ll}\left(y_l + \sum_{k=l+1}^{p} b_{lk} y_k - \mu'_l\right)^2\right\} \quad (7)$$

For the remainder of this paper, we will use m_{ij}, $\mathbf{m}_{i\bullet}$ and $\mathbf{m}_{\bullet j}$ to denote element (i, j), row i and column j of a matrix M respectively.

Factoring the likelihood in this way gives an insight into how the dependencies between dimensions are captured by a full covariance multivariate Gaussian distribution. Element y_l of observation $\mathbf{y}$ is modelled as a linear regression where $y_{l+1}, \ldots, y_p$ are the explanatory variables and $b_{l,l+1}, \ldots, b_{lp}$ are the regression coefficients. Thus y_p is distributed as an unconditional univariate Gaussian $y_p \sim N(\mu'_p, s_{pp})$, element y_{p-1} is conditioned on y_p with $y_{p-1}|y_p \sim N(\mu'_{p-1} - b_{p-1,p}\, y_p,\ s_{p-1,p-1})$ and so on until y_1 which is conditioned on all other elements of $\mathbf{y}$. These observations show how dimensions i and j of $\mathbf{y}$ can be modelled as statistically independent by setting $b_{ij} = 0$ (recall b_{ij} can only be non-zero for $j > i$), and conversely including non-zero b_{ij} incorporates any dependence between dimensions i and j into the covariance matrix.

1.2 Linear dynamic models

The linear dynamic model (LDM), also known as the Kalman filter model, is the model with which this work is concerned. Letting $\mathbf{y}_t$ and $\mathbf{x}_t$ respectively denote p and q dimensioned continuous-valued observation and state vectors at time t, the LDM is described by the following pair of equations:

$$\mathbf{y}_t = H\mathbf{x}_t + \boldsymbol{\epsilon}_t \qquad \boldsymbol{\epsilon}_t \sim N(\mathbf{v}, C) \quad (8)$$

$$\mathbf{x}_t = F\mathbf{x}_{t-1} + \boldsymbol{\eta}_t \qquad \boldsymbol{\eta}_t \sim N(\mathbf{w}, D) \quad (9)$$

and a distribution over the initial state, $\mathbf{x}_1 \sim N(\boldsymbol{\pi}, \Lambda)$. The LDM is a generative model, giving a time-varying multivariate Gaussian distribution over the observations. Underlying dynamics are modelled by the state evolution which is according to a first-order auto-regressive (AR) process. Equation 8 describes the *observation process* and Equation 9 describes the *state process*. The LDM comes from the family of linear Gaussian models (for further information see Roweis and Ghahramani (1999) or Rosti and Gales (2001)), and was first applied to speech recognition by Digalakis (1992), work which has been continued in recent times by Ma and Deng (2004), Rosti (2004) and Frankel and King (2007).

The observation noise $\boldsymbol{\epsilon}_t$ is typically set to have a diagonal covariance matrix

(such as in Digalakis (1992) and Rosti (2004)), in which case the distribution of errors is approximated by a projection of a lower dimensioned state via the observation matrix H. This gives a model with significantly fewer parameters than one with a fully specified noise covariance matrix, though represents a loss in generality (Roweis and Ghahramani, 1999). Results reported in Frankel and King (2007) show that phone classification using LDMs with full noise covariance matrices yields higher accuracy than where diagonal covariances are used. Insight into this finding is offered in the plots of Figure 2, which show the correlation and mutual information structure of both speech parameters and LDM prediction errors. The speech data comprises 480 validation utterances taken from the TIMIT (Lamel et al., 1986) training set. The prediction errors are calculated during a forward Kalman filter pass through the same validation utterances using a set of LDMs with parameters estimated on the remainder of the training data. The model used to filter each segment is chosen according to the time-aligned phone labels.

The top two plots show the data and error correlations, where dark squares signify high correlation between a pair of dimensions. With the 39 dimensioned feature vector comprising a concatenation of 12 MFCCs, energy, and their first and second derivatives, the top left plot shows that correlations tend to be highest among lower order cepstral coefficients and their derivatives. Furthermore, the diagonal line in the upper right of the plot shows that strong correlations exist between features and their second derivatives. The error correlation plot has been normalized by the data variance to show that some of the structure present in the data has been accounted for by the model. An ideal model would explain all the structure in the data, leaving uncorrelated errors, however these plots demonstrate that spatial correlations persist.

The mutual information $I(Y;X)$ gives a measure of how much information one random variable provides about another, and for continuous-valued variables, such as feature dimensions, can be calculated using histogram-based methods (Moddemeijer, 1989). With dark squares corresponding to high values, the lower two plots show the mutual information between each pair of dimensions for the original features and prediction errors. Many of the attributes of the correlation plots are found here also: mutual information tends to be higher between lower ordered cepstral coefficients, and also between features and their second derivatives, as well as between the cepstral coefficients and energy. These plots show that the LDMs have accounted for some, but not all, of the mutual information between feature dimensions.

The LDM includes three Gaussian covariance matrices: those of the observation and state noise distributions, C and D respectively, and the initial state covariance Λ. With adequate training data, fully specified covariance matrices can be estimated. However, practical applications frequently encounter problems of data sparsity, especially if models are context dependent such as

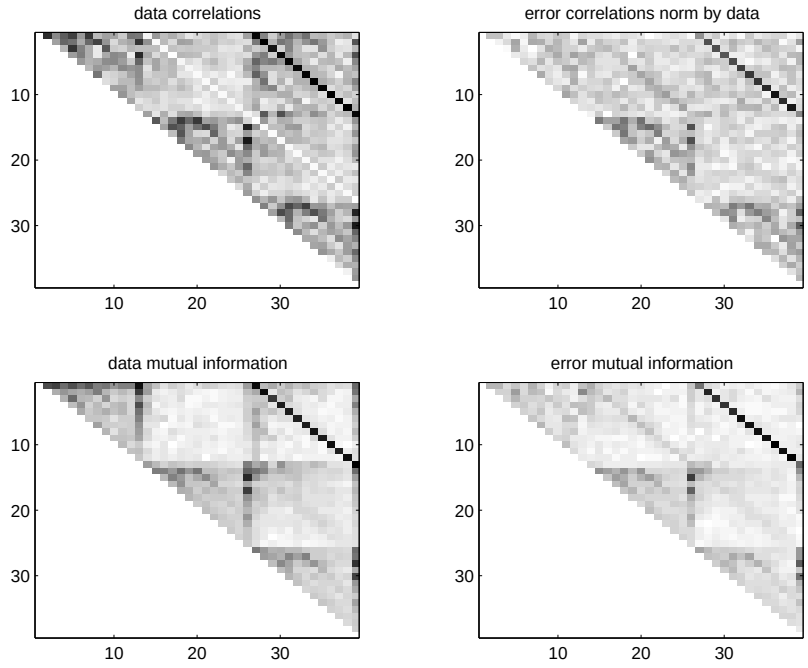

Fig. 2. Graphical representation of the correlation and mutual information structure of the speech parameters and LDM prediction errors from the TIMIT validation set.

in Rosti (2004), or there is switching between multiple models per phone as in Ma and Deng (2004). This work considers modelling the LDM's precision rather than covariance matrices, thereby allowing the factorization outlined above. Such an approach facilitates modelling a subset of the possible error dependencies, producing error covariances which are between diagonal and full, and furthermore allows flexible tying schemes in which the error structure can

be shared between models and model-specific magnitudes estimated. These techniques should prove useful tools in making the best use of available data.

2 Deriving EM updates for factored-covariance LDMs

With $\mathcal{Y} = \mathbf{y}_1^N$ and $\mathcal{X} = \mathbf{x}_1^N$ denoting sequences of N observation and state vectors respectively, the Markovian structure of the model means that the joint likelihood of state and observations can be written as

$$L(\Theta|\mathcal{Y},\mathcal{X}) = f(\mathcal{Y},\mathcal{X}|\Theta) = f(\mathbf{x}_1|\Theta)\prod_{t=2}^{N} f(\mathbf{x}_t|\mathbf{x}_{t-1},\Theta)\prod_{t=1}^{N} f(\mathbf{y}_t|\mathbf{x}_t,\Theta) \tag{10}$$

This yields a log-likelihood function of

$$l(\Theta|\mathcal{Y},\mathcal{X}) = \log f(\mathcal{Y},\mathcal{X}|\Theta) = l_{state}(\Theta|\mathcal{Y},\mathcal{X}) + l_{obs}(\Theta|\mathcal{Y},\mathcal{X}) \tag{11}$$

where $l_{state}(\Theta|\mathcal{Y},\mathcal{X})$ and $l_{obs}(\Theta|\mathcal{Y},\mathcal{X})$ denote the contributions to the log-likelihood of the state and observation respectively:

$$l_{state}(\Theta|\mathcal{Y},\mathcal{X}) = \log f(\mathbf{x}_1|\Theta) + \sum_{t=2}^{N}\log f(\mathbf{x}_t|\mathbf{x}_{t-1},\Theta) \tag{12}$$

$$l_{obs}(\Theta|\mathcal{Y},\mathcal{X}) = \sum_{t=1}^{N}\log f(\mathbf{y}_t|\mathbf{x}_t,\Theta) \tag{13}$$

Given that the state noise covariance can be set to the identity or a diagonal matrix with no loss in generality (Roweis and Ghahramani, 1999), we do not consider factorizations of the state process parameters D and Λ [1]. Noting that the log-likelihood function of Equation 11 is linearly separable in state and observation parameters, we only consider $l_{obs}(\Theta|\mathcal{Y},\mathcal{X})$ in the derivation which follows.

From Equation 8 we can write the pdf of the observation given state as:

$$f(\mathbf{y}_t|\mathbf{x}_t,\Theta) = \frac{1}{\sqrt{(2\pi)^p|C|}}\exp\left\{-\frac{1}{2}(\mathbf{y}_t - H\mathbf{x}_t - \mathbf{v})^T C^{-1}(\mathbf{y}_t - H\mathbf{x}_t - \mathbf{v})\right\} \tag{14}$$

We now make the following substitution:

$$C^{-1} = U_C^T\, S_C\, U_C \tag{15}$$

[1] The techniques presented in this paper readily transfer to the state parameters.

where as above, U_C is unit upper-diagonal and S_C is a diagonal matrices. By Equation 13, and using an analogous rearrangement to that which gave Equation 6, we find that

$$l_{obs}(\Theta|\mathcal{Y},\mathcal{X}) \propto \\ -\frac{1}{2}\sum_{t=1}^{N}\left\{-\log|S_C| + (\mathbf{y}_t + B_C\mathbf{y}_t - H^{'}\mathbf{x}_t - \mathbf{v}^{'})^T S_C(\mathbf{y}_t + B_C\mathbf{y}_t - H^{'}\mathbf{x}_t - \mathbf{v}^{'})\right\} \tag{16}$$

where $U_C = I + B_C$, $H^{'} = U_C H$ and $\mathbf{v}^{'} = U_C\mathbf{v}$.

2.1 Estimation with an observable state

Maximum likelihood (ML) parameter estimation involves finding the parameter set which given some data, maximizes the likelihood (or equivalently, and frequently more simply, the log-likelihood) function. For LDMs, this is complicated by the hidden nature of the state, and so it is useful to first consider a slightly altered scenario where the state is in fact observed. In this case, true ML estimates of the model parameters can be produced by maximizing the log-likelihood function given in Equation 16 for each parameter in turn. This is a question of producing partial derivatives, equating to zero and solving.

In order to ensure an upper-diagonal structure with zeros along the diagonal in the estimate of B_C, maximization must use the log-likelihood function written out explicitly as a set of sums.

It was shown above how a non-zero value of b^c_{ij} (for $j > i$) occurs in a Gaussian which models the dependency between dimensions i and j. To allow the specification during estimation of which dimensions should be modelled as dependent, and which to be assumed statistically independent, we introduce the following notation. Let $\mathcal{B}^c_i$ be an ordered set containing the column indices of the non-zero elements of row i of B_C, and $|B_C|$ denote the number of elements in the set. By definition, this set has the property that $\forall j \in \mathcal{B}^c_i,\ j > i$. Then for a matrix M, $[\{m_{ij}\}_{j\in\mathcal{B}^c_i}]$ denotes a $1 \times |\mathcal{B}^c_i|$ row vector containing the elements from row i of M which appear in the columns indexed by $\mathcal{B}^c_i$.

Using the notation introduced above, we can then use $[\{\mathbf{m}_{k\bullet}\}_{k\in\mathcal{B}^c_i}]$ to denote a $|\mathcal{B}^c_i| \times p$ submatrix consisting of the rows of M which are indexed by the non-zero elements of row i of B_C. By extension, $[\{\mathbf{m}_{kj}\}_{k\in\mathcal{B}^c_i,\, j\in\mathcal{B}^c_i}]$ represents a submatrix of size $|\mathcal{B}^c_i| \times |\mathcal{B}^c_i|$ containing the elements in rows and columns indexed by $\mathcal{B}^c_i$.

Given that for a diagonal matrix such as S_C,

$$|S_C| = \prod_{l=1}^{p} s_{ll}^c \tag{17}$$

we can express the observation process log-likelihood of Equation 16 as the following set of sums:

$$\begin{aligned} &l_{obs}(\Theta|\mathcal{Y},\mathcal{X}) \propto \\ &-\frac{1}{2}\sum_{t=1}^{N}\left\{-\sum_{l=1}^{p}\log s_{ll}^c + \sum_{l=1}^{p} s_{ll}^c\left(y_{lt} + \sum_{k\in\mathcal{B}_l^c} b_{lk}^c y_{kt} - \sum_{k=1}^{q} h_{lk}^{'} x_{kt} - v_l^{'}\right)^2\right\} \end{aligned} \tag{18}$$

Partial derivatives can then be taken in terms of $h_{ij}^{'}$, $v_i^{'}$ and b_{ij}^c, the individual elements of the observation process parameters $H^{'}$, $\mathbf{v}^{'}$ and B_C. This proceeds as follows for $h_{ij}^{'}$:

$$\begin{aligned} \frac{\partial l_{obs}}{\partial h_{ij}^{'}} &= \sum_{t=1}^{N} s_{ii}^c\left(y_{it} + \sum_{k\in\mathcal{B}_i^c} b_{ik}^c y_{kt} - \sum_{k=1}^{q} h_{ik}^{'} x_{kt} - v_i^{'}\right) x_{jt} = 0 \\ \Rightarrow \sum_{k=1}^{q}\hat{h}_{ik}^{'}\sum_{t=1}^{N} x_{kt}x_{jt} &= \sum_{k\in\mathcal{B}_i^c}\hat{b}_{ik}^c\sum_{t=1}^{N} y_{kt}x_{jt} - \hat{v}_i^{'}\sum_{t=1}^{N} x_{jt} + \sum_{t=1}^{N} y_{it}x_{jt} \end{aligned} \tag{19}$$

and for $v_i^{'}$:

$$\begin{aligned} \frac{\partial l_{obs}}{\partial v_i^{'}} &= \sum_{t=1}^{N} s_{ii}^c\left(y_{it} + \sum_{k\in\mathcal{B}_i^c} b_{ik}^c y_{kt} - \sum_{k=1}^{q} h_{ik}^{'} x_{kt} - v_i^{'}\right) = 0 \\ \Rightarrow \hat{v}_i^{'} &= \frac{1}{N}\sum_{t=1}^{N} y_{it} + \frac{1}{N}\sum_{k\in\mathcal{B}_i^c}\hat{b}_{ik}^c\sum_{t=1}^{N} y_{kt} - \frac{1}{N}\sum_{k=1}^{q}\hat{h}_{ik}^{'}\sum_{t=1}^{N} x_{kt} \end{aligned} \tag{20}$$

Finally, taking partial derivatives with respect to b_{ij}^c for $j \in \mathcal{B}_i^c$ (the non-zero elements in row i of B_C), gives:

$$\begin{aligned} \frac{\partial l_{obs}}{\partial b_{ij}^c} &= -\sum_{t=1}^{N} s_{ii}^c\left(y_{it} + \sum_{k\in\mathcal{B}_i^c} b_{ik}^c y_{kt} - \sum_{k=1}^{q} h_{ik}^{'} x_{kt} - v_i^{'}\right) y_{jt} = 0 \\ \Rightarrow \sum_{k\in\mathcal{B}_i^c}\hat{b}_{ik}^c\sum_{t=1}^{N} y_{kt}y_{jt} &= \sum_{k=1}^{q}\hat{h}_{ik}^{'}\sum_{t=1}^{N} x_{kt}y_{jt} + \hat{v}_i^{'}\sum_{t=1}^{N} y_{jt} - \sum_{t=1}^{N} y_{it}y_{jt} \end{aligned} \tag{21}$$

We now introduce the set of sufficient statistics $\Gamma^{(x)}$, $\Gamma^{(y)}$, $\Gamma^{(xx)}$, $\Gamma^{(yy)}$ and $\Gamma^{(yx)}$, defined as:

$$\Gamma^{(x)} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{x}_t \quad \Gamma^{(y)} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t \tag{22}$$

$$\Gamma^{(xx)} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{x}_t\mathbf{x}_t^T \quad \Gamma^{(yy)} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t\mathbf{y}_t^T \tag{23}$$

$$\Gamma^{(yx)} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t\mathbf{x}_t^T \tag{24}$$

Note that we follow the notational convention introduced above, and use $\gamma_{i,j}^{(x)}$ to reference element i, j of $\Gamma^{(x)}$. Then for $i = 1, \ldots, p$, Equations 19, 20, and 21 can be rewritten as a set of $q + 1 + |\mathcal{B}_i^c|$ simultaneous equations in matrix form:

$$\begin{aligned}
\hat{\mathbf{h}}'_{i\bullet}\,\mathbf{\Gamma}^{(xx)} + \hat{v}_i'\,\Gamma^{(x)T} - [\{b_{ij}\}_{j\in\mathcal{B}_i^c}]\,[\{\gamma_{k\bullet}^{(yx)}\}_{k\in\mathcal{B}_i^c}] &= \boldsymbol{\gamma}_{i\bullet}^{(yx)} \\
\hat{\mathbf{h}}'_{i\bullet}\,\Gamma^{(x)} + \hat{v}_i' - [\{b_{ij}\}_{j\in\mathcal{B}_i^c}]\,[\{\gamma_{k}^{(y)}\}_{k\in\mathcal{B}_i^c}] &= \gamma_{i}^{(y)} \\
\hat{\mathbf{h}}'_{i\bullet}\,[\{\gamma_{k\bullet}^{(yx)}\}_{k\in\mathcal{B}_i^c}]^T + \hat{v}_i'\,[\{\gamma_{k}^{(y)}\}_{k\in\mathcal{B}_i^c}]^T - [\{b_{ij}\}_{j\in\mathcal{B}_i^c}]\,[\{\gamma_{kj}^{(yy)}\}_{k\in\mathcal{B}_i^c,\,j\in\mathcal{B}_i^c}] &= [\{\gamma_{ij}^{(yy)}\}_{j\in\mathcal{B}_i^c}]
\end{aligned} \tag{25}$$

With the number of parameters and hence simultaneous equations varying according to the size of $\mathcal{B}_i^c$, closed form parameter estimates must be found for each row in turn. Combining the set of equations in 25 into partitioned matrices yields:

$$\begin{aligned}
&\left[\hat{\mathbf{h}}'_{i\bullet} \;\; v_i' \;\; [\{\hat{b}_{ij}\}_{j\in\mathcal{B}_i^c}]\right] \\
&= \left[\boldsymbol{\gamma}_{i\bullet}^{(yx)} \;\; \gamma_i^{(y)} \;\; [\{\gamma_{ij}^{(yy)}\}_{j\in\mathcal{B}_i^c}]\right]
\begin{bmatrix}
\Gamma^{(xx)} & \Gamma^{(x)} & [\{\gamma_{k\bullet}^{(yx)}\}_{k\in\mathcal{B}_i^c}]^T \\
\Gamma^{(x)T} & 1 & [\{\gamma_{k}^{(y)}\}_{k\in\mathcal{B}_i^c}]^T \\
-[\{\gamma_{k\bullet}^{(yx)}\}_{k\in\mathcal{B}_i^c}] & -[\{\gamma_{k}^{(y)}\}_{k\in\mathcal{B}_i^c}] & -[\{\gamma_{kj}^{(yy)}\}_{k\in\mathcal{B}_i^c,\,j\in\mathcal{B}_i^c}]
\end{bmatrix}^{-1}
\end{aligned} \tag{26}$$

Given the upper-diagonal form of $U_C = B_C + I_p$, inversion to find U_C^{-1} and hence $U_C = U_C^{-1}H'$ and $\mathbf{v} = U_C^{-1}\mathbf{v}'$ is straightforward.

The parameter S_C which accompanies B_C in specifying the observation noise precision must also be maximized. Using the result that for symmetric Z,

$$\frac{d}{dZ}\log|Z| = Z^{-1} \tag{27}$$

$\hat{S}_C$ can be found as follows:

$$\begin{aligned}
\frac{\partial l}{\partial S_C} &= N S_C^{-1} - \sum_{t=1}^{N} (\mathbf{y}_t + B_C \mathbf{y}_t - H^{'} \mathbf{x}_t - \mathbf{v}^{'})(\mathbf{y}_t + B_C \mathbf{y}_t - H^{'} \mathbf{x}_t - \mathbf{v}^{'})^T \\
&= N S_C^{-1} - \sum_{t=1}^{N} \left((I + B_C)\mathbf{y}_t - H^{'} \mathbf{x}_t - \mathbf{v}^{'} \right) \left((I + B_C)\mathbf{y}_t - H^{'} \mathbf{x}_t - \mathbf{v}^{'} \right)^T \\
\Rightarrow \hat{S}_C^{-1} &= \operatorname{diag} \left[\frac{1}{N} (I + \hat{B}_C) \sum_{t=1}^{N} \mathbf{y}_t \left((I + \hat{B}_C)\mathbf{y}_t - \hat{H}' \mathbf{x}_t - \hat{\mathbf{v}}^{'} \right)^T \right. \\
&\quad \left. - \frac{1}{N} \sum_{t=1}^{N} \left(\hat{H}' \mathbf{x}_t + \hat{\mathbf{v}}^{'} \right) \left((I + \hat{B}_C)\mathbf{y}_t - \hat{H}' \mathbf{x}_t - \hat{\mathbf{v}}^{'} \right)^T \right] \qquad (28) \\
&= \operatorname{diag} \left[\hat{U}_C \left(\Gamma^{(yy)} \hat{U}_C^T - \Gamma^{(yx)} H'^T - \Gamma^{(y)} \mathbf{v}'^T \right) - \hat{H}' \Gamma^{(yx)T} \hat{U}_C^T \right. \\
&\quad \left. + \hat{H}' \Gamma^{(xx)} \hat{H}'^T + \hat{H}' \Gamma^{(x)} \hat{\mathbf{v}}'^T - \hat{\mathbf{v}}' \Gamma^{(y)T} \hat{U}_C^T + \hat{\mathbf{v}}' \Gamma^{(x)T} \hat{H}'^T + \hat{\mathbf{v}}' \hat{\mathbf{v}}'^T \right] \qquad (29)
\end{aligned}$$

where we use $\operatorname{diag}(M)$ to denote the diagonal matrix whose diagonal elements are the elements of M.

This expression can be simplified using the matrix form of the ML estimates of $\hat{H}^{'}$ and $\hat{\mathbf{v}}^{'}$. Taking partial derivatives of the log-likelihood function given in 16 and equating to 0 yields:

$$\hat{H}' = \left(\hat{U}_C \Gamma^{(yx)} - \hat{\mathbf{v}}' \Gamma^{(x)T} \right) \Gamma^{(xx)-1} \qquad (30)$$

$$\hat{\mathbf{v}}^{'} = \hat{U}_C \Gamma^{(y)} - \hat{H}' \Gamma^{(x)} \qquad (31)$$

Now substituting these expressions into the 3^{rd} and last terms of Equation 29 respectively, we find that all terms but the first cancel, so that

$$\hat{S}_C^{-1} = \operatorname{diag} \left[U_C \left(\Gamma^{(yy)} - \Gamma^{(yx)} H^T - \Gamma^{(y)} \mathbf{v}^T \right) U_C^T \right] \qquad (32)$$

The estimates of factored covariance observation process parameters given in Equations 26 and 32 simply extend to using multiple time series for estimation, in which case the sufficient statistics are averaged over all relevant data points. Similarly, parameter tying can be implemented by pooling the sufficient statistics required to compute a given parameter among the models between which it will be shared.

2.2 Estimation with state hidden – application of the EM algorithm

The expectation maximization (EM) algorithm (Dempster et al., 1977; Bilmes, 1997) provides a means of iterating toward the ML solution in situations where there is missing or incomplete data. In this case the incomplete data is the state, and EM takes a model with parameters $\Theta^{(i)}$ at the i^{th} iteration and makes an update to give $\Theta^{(i+1)}$, such that the likelihood over the training data is increased or left unchanged.

For distributions from the exponential family (of which the LDM with its Gaussian output distribution is a member), the EM algorithm consists of alternating between computing the complete-data conditional expectations of the standard ML sufficient statistics using the most recent parameter set, and using these expectations to update the parameter estimates (Dempster et al., 1977). Writing $\mathbf{x}_t|\mathcal{Y} \sim N(\hat{\mathbf{x}}_{t|N}, \Sigma_{t|N})$, the expectations which must be computed are given by:

$$E[\mathbf{x}_t|\mathcal{Y}, \Theta^{(i)}] = \hat{\mathbf{x}}_{t|N} \tag{33}$$

$$E[\mathbf{x}_t\mathbf{x}_t^T|\mathcal{Y}, \Theta^{(i)}g] = \Sigma_{t|N} + \hat{\mathbf{x}}_{t|N}\hat{\mathbf{x}}_{t|N}^T \tag{34}$$

$$E[\mathbf{x}_t\mathbf{x}_{t-1}^T|\mathcal{Y}, \Theta^{(i)}] = \Sigma_{t,t-1|N} + \hat{\mathbf{x}}_{t|N}\hat{\mathbf{x}}_{t-1|N}^T \tag{35}$$

A Rauch-Tung-Striebel (RTS) smoother (Rauch, 1963) can be used to compute the complete-data state statistic estimates $\hat{\mathbf{x}}_{t|N}$ and $\Sigma_{t|N}$, with the cross-covariance $\Sigma_{t,t-1|N}$ found using the additional recursion given by Digalakis et al. (1993) and Rosti and Gales (2001).

EM for LDMs therefore consists of evaluating the ML parameter estimates given in Equations 26 and 32, replacing $\mathbf{x}_t$, $\mathbf{x}_t\mathbf{x}_t^T$, and $\mathbf{x}_t\mathbf{x}_{t-1}^T$ with their expectations 33–35.

3 Experiments

Two sets of experiments are described below which illustrate the operation of the proposed method. The first uses synthetic data, and the second uses speech data from the TIMIT corpus. In both cases, the LDMs have diagonal initial state covariance Λ and state error covariance D, and it is the form of the observation noise covariance C which is varied. Models are trained on the data corresponding to each of a number of classes using the EM algorithm as described above. During classification, model likelihoods are computed by making a forward Kalman filter pass through each phone segment (token) as described in Frankel and King (2007). In the synthetic data experiments of

Section 3.1, the model with the highest likelihood is chosen, and in the speech data experiments of Section 3.2 a Viterbi search with a bigram language model is used to choose the most likely sequence of phone models (this is valid since the state is reset at phone boundaries). Frankel and King (2007) contains a full description of the experimental setup.

3.1 Synthetic data

This experiment compares the classification of unseen spatially-correlated time series data using LDMs with three different observation noise models. These are: diagonal covariance, full covariance, and factored precision with the transform component tied across all models. The performance of these three models is compared for varying amounts of training data.

The synthetic data was generated using a set of 18 LDMs with 4-dimensional states. Each of these models had originally been trained on 13-dimensional speech parameters corresponding to one of 18 distinct phone classes (data used came from TIMIT corpus as described in Section 3.2 below). Before using these models to generate a new set of data, the observation noise covariance matrices were modified in the following way: a single B_v matrix was chosen to be used for all models, and the elements within the model-specific diagonal S_v matrices were normalized to have a uniform mean and variance across the 18 models. This step was included to prevent classification of unseen data being simplified by significant variation in magnitude of the observation noise, whilst retaining the structure of variation between dimensions within models.

The set of LDMs were then used to generate various sizes of data sets (see table 1) comprising equal numbers of tokens from each class. Each token varied between 4 and 30 frames in length, with the duration chosen at random from a uniform distribution. Of these tokens, 10% were used as validation data, 20% for testing, and the remaining 70% as training data.

Classification then proceeds as follows: an LDM is trained on the data corresponding to each of the 18 classes and the parameters of the model set are stored after each training iteration. The model set which gives the highest classification accuracy on the validation set is then used for classification of the test data. Table 1 and Figure 3 show the classification accuracy for LDMs with the three different observation noise models as the amount of training data was varied from 2 to 50 tokens per class. Modelling the spatial correlations in the observation noise distributions gives very large accuracy increases over those of the diagonal covariance models, regardless of the amount of available training data, and gives improvements over full covariance models, for small data set sizes.

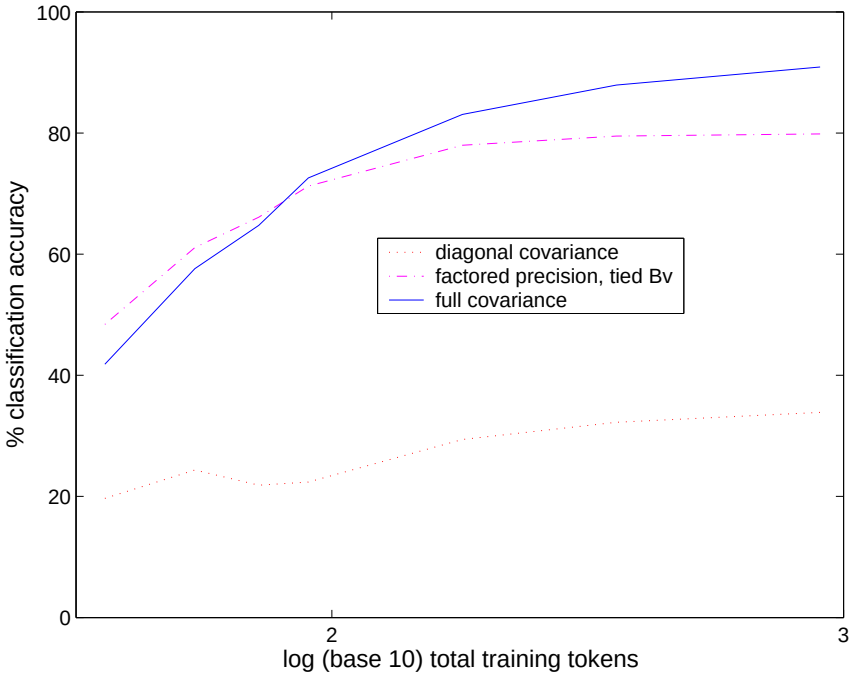

Fig. 3. Classification accuracy on synthetic data as a function of quantity of training data with three different observation noise models: diagonal covariance, full covariance and factored precision with the transform component tied across all models.

3.2 Speech data

This experiment is similar to the first and examines the effect of varying the form of observation noise model with parameters estimated on a range of train-

# training tokens	Classification accuracy		
	diagonal	tied B_v	full
36	19.7%	**48.4%**	41.8%
54	24.4%	**61.1%**	57.6%
72	21.8%	**66.1%**	64.8%
90	22.4%	71.2%	**72.6%**
180	29.4%	78.0%	**83.1%**
360	32.2%	79.5%	**87.9%**
900	33.9%	79.8%	**90.9%**
# parameters	1980	2058	3384

Table 1
Classification accuracy on synthetic data using LDMs with three different observation noise models: diagonal covariance, full covariance, and factored precision with the transform component tied across all models. Results are given with the amount of training data varying from 2 to 50 tokens per class (18 classes). For each separate training set, the highest accuracy is shown in bold face. These results are shown graphically in Figure 3.

ing set sizes, but this time using speech data. The classification procedure is as above, except that after the validation data has been used to determine the number of training iterations, models are retrained on data from the combined training and validation sets prior to evaluation on the test data.

In addition to diagonal, tied-transform and full models, partially specified observation noise models are now considered. The parameter estimation of Section 2 shows how to estimate noise covariance matrices which are between diagonal and full by including a subset of the possible spatial dependencies. The dependencies to include were chosen for this experiment in the following way: a set of diagonal covariance LDMs was trained, and prediction errors calculated on held out validation data as described in Section 1.2. The mutual information between dimensions of the prediction errors was then calculated and a new set of LDMs trained in which 0.5%, 1%, 5%, 10%, 25%, 50% or 75% of spatial dependencies were included, with ranked mutual information used to determine the order in which dependencies were included. Using 0% and 100% of dependencies corresponds to diagonal and full covariance LDMs with 528 and 1269 free parameters respectively. For each size of training set, the models giving the highest phone classification accuracy on the validation data were then chosen for final evaluation on the test data. This process was repeated for each size of training set, with a set of dependencies chosen for each phone.

The results presented in Table 2 and Figure 4 show that where data is limited,

# training tokens	Classification accuracy				
	diagonal	partial		tied B_v	full
305	24.9%	**41.4%**	(0.5%)	30.3%	12.7%
610	39.5%	**45.0%**	(0.5%)	37.0%	21.7%
1220	**49.5%**	**49.5%**	(0.5%)	40.1%	32.7%
3025	52.4%	**53.0%**	(1%)	47.6%	48.7%
6025	**58.1%**	**58.1%**	(1%)	54.6%	56.5%
11896	63.5%	64.2%	(75%)	61.7%	**64.4%**
28954	67.8%	69.5%	(75%)	66.7%	**69.9%**
142780	68.9%	71.4%	(75%)	68.9%	**72.0%**

Table 2
Classification accuracy using LDMs on speech data with four different observation noise models: diagonal covariance, partially-specified (% possible dependencies given in parentheses), factored precision with the transform component tied across all models, and full covariance. Results are given with the number of training tokens varying from 305 to 142780 (full TIMIT train set). For each separate training set, the highest accuracy is shown in bold face. Figure 4 presents these results graphically.

partially specified covariance models give the highest classification accuracies. Given sufficient data, full covariance models give the best performance, but partially-specified covariance models are very close behind.

For the 5 smallest training sets, the gains from using partially specified covariance models over full covariance models are statistically significant (consistent across the test data). For two of the eight training sets, partially specified and diagonal models give equal classification accuracy, and for the remainder the partially specified models yield statistically significant accuracy increases over the diagonal models.

4 Conclusions

This work has considered the form of the observation noise model for LDMs and introduces a technique for estimating factored precision matrices. This allows estimation of covariance matrices other than simply diagonal or full, and facilitates separate tying of the transform and magnitude components between models.

In the illustrative experiments, the partially specified covariance models are as good as diagonal covariance models when there are few data and as good as full covariance models with larger amounts of data. Furthermore, the number

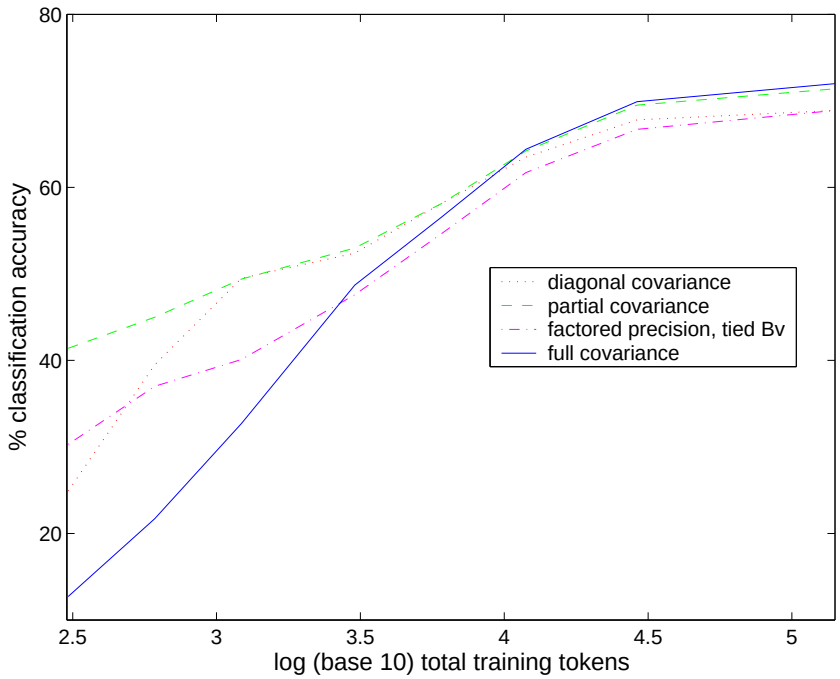

Fig. 4. Classification accuracy as a function of the quantity of training data using LDMs on TIMIT speech data with four different observation noise models: diagonal covariance, partial set of dependencies, factored precision with the tied transform, and full covariance. Note that, for all data set size, the partially-specified covariance model gives either the highest, or very close to the highest accuracy.

of model parameters can be varied smoothly between those two extremes, to adapt to any size data set.

Given sufficient training data, these results support the use of full covariance observation noise models [2]. However, the techniques introduced in this paper have been shown to make efficient use of small training sets. As noted above, data sparsity can arise as a result of a particular implementation. For example, switching models require multiple LDMs per phone (Ma and Deng, 2004), or the data sparsity inherent in a system based on triphones (context-dependent phone models) (Rosti, 2004) makes tying at some level inevitable. Modelling factored precision matrices should prove a useful tool in such cases. Furthermore, in a implementation such as triphones, it may be reasonable to assume that within clusters, the errors would be distributed in a similar enough fashion that tying transforms would come into its own.

Acknowledgements

Many thanks to Stephen Isard for helpful discussions whilst this work was in preparation.

References

Bilmes, J., 1997. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Tech. Rep. ICSI-TR-97-021, University of Berkeley.

Bilmes, J., 2000. Factored sparse inverse covariance matrices. In: Proc. ICASSP.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B (39), 1–38.

Digalakis, V., 1992. Segment-based stochastic models of spectral dynamics for continuous speech recognition. Ph.D. thesis, Boston University Graduate School.

Digalakis, V., Rohlicek, J., Ostendorf, M., October 1993. ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. IEEE Trans. Speech and Audio Processing 1 (4), 431–442.

Frankel, J., King, S., January 2007. Speech recognition using linear dynamic models. IEEE Transactions on Speech and Audio Processing 15 (1), 246–256.

Gales, M., May 1999. Semi-tied covariance matrices for hidden Markov models. IEEE Transactions on Speech and Audio Processing 7 (3), 272–281.

[2] Fully specified observation noise covariances only cause a marginal increase in computation as the Kalman filter recursions yield full prediction error covariance matrices.

Gold, B., Morgan, N., 1999. Speech and Audio Signal Processing. Wiley Press.
Lamel, L., Kassel, R., Seneff, S., February 1986. Speech database development: design and analysis of the acoustic-phonetic corpus. In: Proc. Speech Recognition Workshop. Palo Alto, CA., pp. 100–109.
Ljolje, A., 1994. The importance of cepstral parameter correlations in speech recognition. Computer Speech and Language 8, 223–232.
Ma, J., Deng, L., 2004. A mixed-level switching dynamic system for continuous speech recognition. Computer Speech and Language 18, 49–65.
Macho, D., Nadeu, C., Jancovic, P. Rozinaj, G., Hernando, J., 1999. Comparison of time and frequency filtering and cepstral-time matrix approaches in ASR. In: Proc. Eurospeech. Budapest, Hungary, pp. 77–80.
Moddemeijer, R., 1989. On estimation of entropy and mutual information of continuous distributions. Signal Processing 16 (3), 233–246.
Olsen, P., Gopinath, R., January 2004. Modeling inverse covariance matrices by basis expansion. IEEE Transactions on Speech and Audio Processing 12 (1), 37–46.
Rauch, H. E., 1963. Solutions to the linear smoothing problem. IEEE Transactions on Automatic Control 8, 371–372.
Rosti, A., Gales, M., 2001. Generalised linear Gaussian models. Tech. Rep. CUED/F-INFENG/TR.420, Cambridge University Engineering.
Rosti, A.-V., 2004. Linear gaussian models for speech recognition. Ph.D. thesis, Machine Intelligence Laboratory, University of Cambridge.
Roweis, S., Ghahramani, Z., 1999. A unifying review of linear Gaussian models. Neural Computation 11 (2).
Young, S., December 1995. Large vocabulary continuous speech recognition: A review. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding. Snowbird, Utah, pp. 3–28.