Contextually Supervised Source Separation with Application to Energy Disaggregation

Matt Wytock and J. Zico Kolter
Machine Learning Department
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

We propose a new framework for single-channel source separation that lies between the fully supervised and unsupervised setting. Instead of supervision, we provide input features for each source signal and use convex methods to estimate the correlations between these features and the unobserved signal decomposition. Contextually supervised source separation is a natural fit for domains with large amounts of data but no explicit supervision; our motivating application is energy disaggregation of hourly smart meter data (the separation of whole-home power signals into different energy uses). Here contextual supervision allows us to provide itemized energy usage for thousands homes, a task previously impossible due to the need for specialized data collection hardware. On smaller datasets which include labels, we demonstrate that contextual supervision improves significantly over a reasonable baseline and existing unsupervised methods for source separation. Finally, we analyze the case of $\ell_2$ loss theoretically and show that recovery of the signal components depends only on cross-correlation between features for different signals, not on correlations between features for the same signal.

Introduction

We consider the single-channel source separation problem, in which we wish to separate a single aggregate signal into a mixture of unobserved component signals. Traditionally, this problem has been approached in two ways: the supervised setting (Kolter, Batra, and Ng 2010; Roweis 2001; Schmidt and Olsson 2006), where we have access to training data with the true signal separations and the unsupervised (or “blind”) setting (Blumensath and Davies 2005; Davies and James 2007; Lewicki and Sejnowski 2000; Schmidt and Morup 2006), where we have only the aggregate signal. However, both settings have potential drawbacks: for many problems, including energy disaggregation—which looks to separate individual energy uses from a whole-home power signal (Hart 1992)—it can be difficult to obtain training data with the true separated signals needed for the supervised setting; in contrast, the unsupervised setting is an ill-defined problem with arbitrarily many solutions, and thus algorithms are highly task-dependent.

In this work, we propose an alternative approach that lies between these two extremes: with contextual supervision, along with the input signal to be separated, we provide contextual features correlated with the unobserved component signals. In practice, we find that this is often much easier than providing a fully supervised training set, yet it also allows for a well-defined problem, unlike the unsupervised setting. The approach is a natural fit for energy disaggregation, since we have strong correlations between energy usage and easily observed context—air conditioning spikes in hot summer months, lighting increases when there is a lack of sunlight, etc. We formulate our model directly as an optimization problem in which we jointly estimate these correlations along with the most likely source separation. Theoretically, we show that when the contextual features are relatively uncorrelated between different groups, we can recover the correct separation with high probability.

Contextually supervised source separation provides a compelling framework for energy disaggregation from “smart meters”, communication-enabled power meters that are currently installed in more than 32 million homes (Institute for Electric Efficiency 2012), but are limited to recording whole home energy usage at low frequencies (every 15 minutes or hour). This is an important task since many studies have shown that consumers naturally adopt energy conserving behaviors when presented with a breakdown of their energy usage (Darby 2006; Neenan and Robinson 2009; Ehrhardt-Martinez, Donnelly, and Laitner 2010). There are several possible ways that such a breakdown could be achieved; for example, by installing current sensors on each device we could monitor electricity use directly. But, as custom hardware installation is relatively expensive (and requires initiative from homeowners), algorithmic approaches that allow disaggregation of energy data already being collected are appealing. However, existing energy disaggregation approaches virtually all use high-frequency sampling (e.g. per second or faster) which still requires the installation of custom monitoring hardware for data collection. In contrast, by enabling disaggregation of readily available low-resolution smart meter data, we can immediately realize the benefits of observing itemized energy use without the need for additional monitoring hardware.
Contextually supervised source separation
We begin by formulating the optimization problem for contextual source separation. Formally, we assume there is some unknown matrix of \( k \) component signals
\[
Y \in \mathbb{R}^{T \times k} = \begin{bmatrix} y_1 & y_2 & \cdots & y_k \end{bmatrix}
\]
from which we observe the sum \( \bar{y} = \sum_{i=1}^{k} y_i \). For example, in our disaggregation setting, \( y_i \in \mathbb{R}^T \) could denote a power trace (with \( T \) total readings) for a single type of appliance, such as the air conditioning, lighting, or electronics, and \( \bar{y} \) denotes the sum of all these power signals, which we observe from a home’s power meter.

In our proposed model, we represent each individual component signal \( y_i \) as a linear function of some component-specific bases \( X_i \in \mathbb{R}^{T \times n_i} \)
\[
y_i \approx X_i \theta_i
\]
where \( \theta_i \in \mathbb{R}^{n_i} \) are the signal’s coefficients. The formal objective of our algorithm is: given the aggregate signal \( \bar{y} \) and the component features \( X_i, i = 1, \ldots, k \), estimate both the parameters \( \theta_i \) and the unknown source components \( y_i \).
We cast this as an optimization problem
\[
\begin{align*}
\text{minimize}_{Y, \theta} & \quad \sum_{i=1}^{k} \{ \ell_i(y_i, X_i \theta_i) + g_i(y_i) + h_i(\theta_i) \} \\
\text{subject to} & \quad \sum_{i=1}^{k} y_i = \bar{y}
\end{align*}
\]
where \( \ell_i : \mathbb{R}^T \times \mathbb{R}^T \to \mathbb{R} \) is a loss function penalizing differences between the \( i \)th reconstructed signal and its linear representation; \( g_i \) is a regularization term encoding the “likely” form of the signal \( y_i \), independent of the features; and \( h_i \) is a regularization penalty on \( \theta_i \). Choosing \( \ell_i, g_i \) and \( h_i \) to be convex functions results in a convex optimization problem.

A natural choice of loss function \( \ell_i \) is a norm penalizing the difference between the reconstructed signal and its features \( \|y_i - X_i \theta_i\| \), but since our formulation enables loss functions that depend simultaneously on all \( T \) values of the signal, we allow for more complex choices as well. For example in the energy disaggregation problem, air conditioning is correlated with high temperature but does not respond to outside temperature changes instantaneously; thermal mass and the varying occupancy in buildings often results in air conditioning usage that correlates with high temperature over some window (for instance, if no one is in a room during a period of high temperature, we may not use electricity then, but need to “make up” for this later when someone does enter the room). In this case, the loss function
\[
\ell_i(y_i, X_i \theta_i) = \|(y_i - X_i \theta_i)(I \otimes 1^T)\|_2^2
\]
which penalizes the aggregate difference of \( y_i \) and \( X_i \theta_i \) over a sliding window, can be used to capture such dynamics. In many settings, it may also make sense to use \( \ell_1 \) or \( \ell_\infty \) rather than \( \ell_2 \) loss, depending on the nature of the source signal.
Likewise, since the objective term $g_i$ depends on all $T$ values of $y_i$, we can use it to encode the likely dynamics of the source signal independent of $X_i\theta_i$. For air conditioning and other single appliance types, we expect sharp transitions between on/off states which we can encode by penalizing the $\ell_1$ norm of $Dy_i$ where $D$ is the linear difference operator subtracting $(y_i)_{t-1} - (y_i)_t$. For other types of energy consumption, for example groups of many electronic appliances, we expect the signal to have smoother dynamics and thus $\ell_2$ loss is more appropriate. Finally, we also include $h_i$ for statistical regularization purposes—but for problems where $T \gg n_i$, such as the ones we consider in energy disaggregation, the choice of $h_i$ is less important.

**Theoretical analysis**

Next, we consider the ability of our model to recover the true source signals as $T$ grows while $k$ and $n_i$ remain fixed. For the purposes of this section only, we restrict our attention to the choice of $\ell_2$ loss, no $g_i$ or $h_i$ terms, and Gaussian noise (the extension to the sub-Gaussian case is straightforward). We show that under this specialization of the model, the optimization problem recovers the underlying signals at a rate dependent on the linear independence between blocks of input features $X_i$. In practice, the choice of $\ell_2$, $g_i$ and $h_i$ is problem-specific and as we see in our experiments, this choice has a significant impact on performance. As we show in this section, for the special case of $\ell_2$ loss with no regularization, the estimate of the $\theta_i$ reduces to the least-squares estimate which simplifies theoretical analysis significantly. However, while this provides intuition on the essential behavior of the model in the large $T$ regime, we note that this is a special case of the framework and that in general more sophisticated loss functions will result in more complex algorithms with better performance.

Formally, for this section we assume the source signals have Gaussian noise

$$y_i = X_i\theta_i^* + w_i$$

for some $\theta_i^* \in \mathbb{R}^{n_i}$ and $w_i \sim \mathcal{N}(0, \sigma_i^2 I)$. Under the choice of $\ell_2$ loss, our optimization problem becomes

$$\min_{\mathcal{Y}, \mathcal{\theta}} \sum_{i=1}^{k} \frac{1}{2} \|y_i - X_i\theta_i\|^2_2$$

subject to $\sum_{i=1}^{k} y_i = \bar{y}$

where $\theta \in \mathbb{R}^n$ is a concatenation of all the $\theta_i$’s, $X \in \mathbb{R}^{T \times n}$ is a concatenation of all the $X_i$’s and $n = \sum_{i=1}^{k} n_i$ is the total number of features. This system of equations has the solution

$$\hat{\theta} = (X^TX)^{-1}X^T\bar{y}$$

which is simply the least-squares solution found by regressing $\bar{y}$ on all of the features $X$.

Since each $y_i$ has its own noise term, we can never expect to recover $y_i$ exactly, but we can recover the true $\theta_i^*$ with analysis that is the same as for standard linear regression. However, in the context of source separation, we are interested in the recovery of the “noiseless” $y_i X_i\theta_i^*$, as this corresponds to the recovery of the underlying signals. This can be a significant advantage as it is often much easier to recover the product of $X_i\theta_i^*$ than the individual $\theta_i^*$ parameters themselves. In particular, as we show in our analysis, accurate signal recovery depends only on the degree to which features in different groups are correlated, not on the correlations contained within a particular group. Concretely, our analysis considers how the root mean squared error

$$\text{RMSE}(X_i\hat{\theta}_i) = \sqrt{\frac{1}{T} \|X_i\hat{\theta}_i - X_i\theta_i^*\|^2_2}$$

vanishes for large $T$.

**Theorem 1.** Given data generated by the model (5), and estimating $\theta$ via (9), we have that

$$\mathbb{E} \left[ \|X_i\hat{\theta}_i - X_i\theta_i^*\|^2_2 \right] = \sigma^2 \text{tr} X_i^TX_i(X^TX_i)^{-1} \leq \sigma^2 n_i \rho_i$$

where $\sigma^2 = \sum_{i=1}^{k} \sigma_i^2$ and $\rho_i = \lambda_{\text{max}}(X_i^TX_i(X_i^TX_i)^{-1})$. Furthermore, for $\delta \leq 0.1$, with probability greater than $1 - \delta$

$$\text{RMSE}(X_i\hat{\theta}_i) \leq \sqrt{\frac{4\sigma^2 n_i \rho_i \log(1/\delta)}{T}}.$$

A key quantity in this theorem is the matrix $X_i^TX_i(X_i^TX_i)^{-1} \in \mathbb{R}^{n_i \times n_i}$; $(X_i^TX_i)^{-1}$ denotes the $i, j$ block of the full inverse $(X^TX)^{-1}$ (i.e., first inverting the joint covariance matrix of all the features, and then taking the $i, j$ block), and this term provides a measure of the linear independence between features corresponding to different signals. To see this, note that if features across different signals are orthogonal, $X^TX$ is block diagonal, and thus $X_i^TX_i(X_i^TX_i)^{-1} = X_i^TX_i(X_i^TX_i)^{-1} = I$, so $\rho_i = 1$. Alternatively, if two features provided for different signals are highly correlated, entries of $(X_i^TX_i)^{-1}$ will have large magnitude that is not canceled by $X_i^TX_i$ and $\rho_i$ will be large. This formalizes an intuitive notion for contextually supervised source separation: for recovering the underlying signals, it does not matter if two features for the same signal are highly correlated (this contrasts to the case of recovering $\theta_i^*$ itself which depends on all correlations), but two correlated signals for different features make estimation difficult; intuitively, if two very similar features are provided for two different source signals, attribution becomes difficult. A particularly useful property of these bounds is that all terms can be computed using just $X_i$, so we can estimate recovery rates when choosing our design matrix.
The proof of Theorem 1 proceeds in two steps. First, using rules for linear transformations of Gaussian random variables, we show that the quantity $X_i(\hat{\theta}_i - \theta^*_i)$ is also (zero-mean) Gaussian, which immediately leads to (11). Second, we derive a tail bound on the probability that $X_i(\hat{\theta}_i - \theta^*_i)$ exceeds some threshold, which leads to the sample complexity bound (12); because this quantity has a singular covariance matrix, this requires a slightly specialized probability bound, given by the following lemma

**Lemma 1.** Suppose $x \in \mathbb{R}^r \sim \mathcal{N}(0, \Sigma)$ with $\text{rank}(\Sigma) = n$. Then

$$P (\|x\|_2^2 \geq t) \leq \left(\frac{t}{n\lambda}\right)^{n/2} \exp \left\{-\frac{1}{2}(t/\lambda - n)\right\} \quad (13)$$

where $\lambda$ is the largest eigenvalue of $\Sigma$.

The proof, along with the complete proof of Theorem 1, is deferred to Appendix A.

### Experimental results

In this section we evaluate contextual supervision for energy disaggregation on one synthetic dataset and two real datasets. On synthetic data we demonstrate that contextual supervision significantly outperforms existing methods (e.g. nonnegative sparse coding) and that by tailoring the loss functions to the expected form of the component signals (as is a feature of our optimization framework), we can significantly increase performance. On real data, we begin with a dataset from Pecan Street, Inc. (http://www.pecanstreet.org/) that is relatively small (less than 100 homes), but comes with labeled data allowing us to validate our unsupervised algorithm quantitatively. Here we show that our unsupervised model does remarkably well in disaggregating sources of energy consumption and improves significantly over a reasonable baseline. Finally, we apply the same methodology to disaggregate large-scale smart meter data from Pacific Gas and Electric (PG&E) consisting of over 4000 homes and compare the results of our contextually supervised model to aggregate statistics from survey data.

In all experiments, we tune the model using hyperparameters that weigh the terms in the optimization objective; in the case of energy disaggregation, the model including hyperparameters $\alpha$ and $\beta$ is shown in Table 3. We set these hyperparameters using a priori knowledge about the relative frequency of each signal over the entire dataset. For energy disaggregation, it is reasonable to assume that this knowledge is available either from survey data (e.g. (U.S. Energy Information Administration 2009)), or from a small number of homes with more fine-grained monitoring, as is the case for the Pecan Street dataset. In both cases, we use the same hyperparameters for all homes in the dataset.

**Disaggregation of synthetic data.** The first set of experiments considers a synthetic generation process that mimics signals that we encounter in energy disaggregation. The process described visually in Figure 1 (top) begins with two signals, the first is smoothly varying over time while the other is a repeating step function

$$X_1(t) = \sin(2\pi t/\tau_1) + 1, \quad X_2(t) = I(t \mod \tau_2 < \tau_2/2) \quad (14)$$

where $I(\cdot)$ is the indicator function and $\tau_1, \tau_2$ are the period of each signal. We also use two different noise models: for the smooth signal we sample Gaussian noise from $\mathcal{N}(0, \sigma^2)$ while for the step function, we sample a distribution with a point mass at zero, uniform probability over $[-1, 0] \cup (0, 1]$ and correlate it across time by summing over a window of size $\beta$. Finally, we constrain both noisy signals to be nonnegative and sum them to generate our input.

We generate data under this model for $T = 50000$ time points and consider increasingly specialized optimization objectives while measuring the error in recovering $Y^* = XD(\theta^*) + W$, the underlying source signals corrupted by

![Figure 1: Synthetic data generation process starting with two underlying signals (top left), corrupted by different noise models (top right), summed to give the observed input (row 2) and disaggregated (rows 3 and 4).](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean prediction</td>
<td>0.3776</td>
</tr>
<tr>
<td>Nonnegative sparse coding</td>
<td>0.2843</td>
</tr>
<tr>
<td>$\ell_2$ loss for $y_1$</td>
<td>0.1205</td>
</tr>
<tr>
<td>$\ell_2$ loss for $y_1$, $\ell_1$ loss for $y_2$</td>
<td>0.0994</td>
</tr>
<tr>
<td>$\ell_2$ loss for $y_1$, $\ell_1$ loss for $y_2$, penalty on $|Dy_i|$</td>
<td>0.0758</td>
</tr>
</tbody>
</table>

Table 1: Performance on disaggregation of synthetic data.

489
noise. As can be seen in Table 1, by using $\ell_1$ loss for $y_2$ and adding $g_i(y_i)$ terms penalizing $\|Dy_1\|_2^2$ and $\|Dy_2\|_1$, error decreases by 37% over just $\ell_2$ loss alone; in Figure 1, we observe that our estimation recovers the true source signals closely with the $g_i$ terms helping to capture the dynamics of the noise model for $w_2$.

As a baseline for this result, we compare to the mean prediction heuristic (predicting at each time point a breakdown proportional to the overall probability of each signal) and to a state-of-the-art unsupervised method, nonnegative sparse coding (Hoyer 2002). We apply sparse coding by segmenting the input signal into 1000 examples of 50 time points (1/4 the period of the sine wave, $X_1(t)$) and fit a sparse model of 200 basis functions. We report the best possible source separation by assigning each basis function according to an oracle measuring correlation with the true source signal and using the best value over a grid of hyperparameters. As can be seen in Table 1, the mean prediction heuristic is nearly 5 times worse and sparse coding is nearly 4 times worse than our best contextually supervised model.

**Energy disaggregation with ground truth.** Next we consider the ability of contextual supervision to recover the sources of energy consumption on a real dataset from Pecan Street consisting of 84 homes each with at least 1 year worth of energy usage data. As contextual information we construct a temperature time series using data from Weather Underground (http://www.wunderground.com/) measuring the temperature at the nearby airport in Austin, Texas. The Pecan Street dataset includes fine-grained energy usage information at the minute level for the entire home with an energy breakdown labeled according to each electrical circuit in the home. We group the circuits into categories representing air conditioning, large appliances and base load and aggregate the data on an hourly basis to mimic the scenario presented by smart meter data.

The specification of our energy disaggregation model is given in Table 3—we capture the non-linear dependence on temperature with radial-basis functions (RBFs), include a “Base” category which models energy used as a function of time of day, and featureless “Appliance” category representing large spikes of energy which do not correspond to any available context. For simplicity, we penalize each category’s deviations from the model using $\ell_1$ loss; but for heating and cooling we first multiply by a smoothing matrix $S_n$ (1’s on the diagonal and $n$ super diagonals) capturing the thermal mass inherent in heating and cooling: we expect energy usage to correlate with temperature over a window of time, not immediately. We use $g_i(y_i)$ and the difference operator to encode our intuition of how energy consumption in each category evolves over time. The “Base” category represents an aggregation of many sources which we expect to evolve smoothly over time, while the on/off behavior in other categories is best represented by the $\ell_1$ penalty. Finally we note that in the Pecan Street data, there is no labeled circuit corresponding exclusively to electric heating (“Heating”), and thus we exclude this category for this dataset.

In Table 2, we compare the results of contextual supervision with the mean prediction heuristic and see that contextual supervision improves by 26% over this baseline which is already better than nonnegative sparse coding. Qualitatively we consider the disaggregated energy results for a sin-
features

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
<th>$f_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>Hour of day</td>
<td>$\alpha_1 | y_1 - X_1 \theta_1 |_1$</td>
</tr>
<tr>
<td>Heating</td>
<td>RBFs over temperatures $&lt; 50^\circ F$</td>
<td>$\alpha_2 | S_2(y_1 - X_3 \theta_2) |_1$</td>
</tr>
<tr>
<td>A/C</td>
<td>RBFs over temperatures $&gt; 70^\circ F$</td>
<td>$\alpha_3 | S_3(y_2 - X_2 \theta_2) |_1$</td>
</tr>
<tr>
<td>Appliance</td>
<td>None</td>
<td>$\alpha_4 | y_4 |_1$</td>
</tr>
</tbody>
</table>

Table 3: Model specification for contextually supervised energy disaggregation.

Figure 4: Disaggregated energy usage for a single home near Fresno, California over a summer week (top left) and a winter week (top right); aggregated over 4000+ homes over nearly four years (bottom)

gle home over one week in Figure 2 and see that contextual supervision correctly assigns energy usage to categories—a large amount of energy is assigned to A/C which cycles on and off roughly corresponding to external temperature, large spikes from appliances happen at seemingly random times and the smoothly varying base load is captured correctly. In Figure 3, we consider the disaggregation results for the same home across the entire time period and see that the contextually supervised estimates correspond very closely to the actual sources of energy consumption.

**Large-scale energy disaggregation.** Next, we turn to the motivating problem for our model: disaggregating large-scale low-resolution smart meter data into its component sources of consumption. Our dataset consists of over 4000 homes and was collected by PG&E from customers in Northern California who had smart meters between 1/2/2008 and 12/31/2011. According to estimations based on survey data, heating and cooling (air conditioning and refrigerators) comprise over 39% of total consumer electricity usage (U.S. Energy Information Administration 2009) and thus are dominant uses for consumers. Clearly, we expect temperature to have a strong correlation with these uses and thus we provide contextual supervision in the form of temperature information. The PG&E data is anonymized, but the location of individual customers is identified at the census block level and we use this information to construct a parallel temperature dataset as in the previous example.

We present the result of our model at two time scales, starting with Figure 4 (top), where we show disaggregated energy usage for a single home over a typical summer and winter week. Here we see that in summer, the dominant source of energy consumption is estimated to be air conditioning due to the context provided by high temperature. In winter, this disappears and is replaced to a smaller extent by heating. In Figure 4 (bottom), itemized energy consumption aggregated across all 4000+ homes demonstrates these basic trends in energy usage. Quantitatively, our model assigns 15.6% of energy consumption to air conditioning and 7.7% to heating, reasonably close to estimations based on survey data (U.S. Energy Information Administration 2009) (10.4% for air conditioning and 5.4% for space heating). We speculate that our higher estimation may be due to the model conflating other temperature-related energy usages (e.g. refrigerators and water heating) or to differences in populations between the survey and smart meter customers.

**Conclusion and discussion**

The disaggregation of smart meter data into itemized energy uses creates large opportunities for increases in efficiency; as smart meters are already widely deployed and have been collecting data for the past several years, millions of homes stand to benefit. However, disaggregating smart meter data is a challenging task due to its low-resolution sampling and lack of supervised information. We believe that with the development of contextual supervision described in this paper, we have made a significant advancement in this area that has been previously dominated by methods that rely on either high-resolution or supervised data that, unlike the smart meter data, is not readily available.

An interesting direction for future work is the explicit connection of our large-scale low-resolution methods with the more sophisticated appliance models developed on smaller supervised datasets with high-frequency measurements. However, there are clear limitations as to what can be observed in a whole home power trace that is only sampled once an hour. The development of refined statistical models that produce confidence intervals around their estimations is one avenue for dealing with this uncertainty. Still, the largest gains are likely to come from an increase in sampling frequency, perhaps in a hybrid approach that varies the sampling rate in order to capture more accurate high-frequency snapshots during periods of high activity.
References


Kim, H.; Marwah, M.; Arlitt, M. F.; Lyon, G.; and Han, J. 2011. Unsupervised disaggregation of low frequency power measurements. In *SDM*, volume 11, 747–758. SIAM.


